Mental Models in Human-AI Interaction: Systematic Review of Empirical Methodologies and Guidelines

Téo Sanchez, Oleksandra Vereschak, Ophelia Deroy

▶ **To cite this version:**

Téo Sanchez, Oleksandra Vereschak, Ophelia Deroy. Mental Models in Human-AI Interaction: Systematic Review of Empirical Methodologies and Guidelines. IUI '26: 31st International Conference on Intelligent User Interfaces, Mar 2026, Paphos Cyprus, Cyprus. pp.663-682, ⟨10.1145/3742413.3789223⟩. ⟨hal-05544238⟩

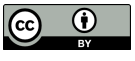
HAL Id: hal-05544238

https://hal.science/hal-05544238v1

Submitted on 9 Mar 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Mental Models in Human-AI Interaction: Systematic Review of Empirical Methodologies and Guidelines

Téo Sanchez
Ludwig Maximilian University
Munich, Germany
teo.sanchez@lmu.de

Oleksandra Vereschak
Independent researcher
Paris, France
olvereschak@gmail.com

Ophelia Deroy
Ludwig Maximilian University
Munich, Germany
ophelia.deroy@lmu.de

Abstract

The notion of *mental model* has long been used in HCI to capture people's understanding and reasoning about computing systems. Eliciting users' mental models can explain their behaviors and attitudes toward a system—why and how they use, rely on, trust, or reject it. However, its use remains conceptually fragmented and methodologically diverse and has not been revisited in light of modern AI systems, whose opacity and newfound abilities may challenge human understanding. To address this gap, we systematically review 88 empirical studies that elicit humans' mental models of AI systems. We extracted and analyzed how studies define and elicit mental models, the type of mental model their method presupposes, and how these vary across AI system types. Drawing from the mental model's framing in cognitive psychology and HCI, and based on descriptive and relational analysis between the variables extracted, we find that (1) mental model elicitations' goal bifurcates between system-specific evaluation and class-level probes surfacing lay theories; (2) epistemic assumptions exceed the classic functional-structural lens (how the system behaves / how it works internally) with analogical and anthropomorphic framings of AI systems; (3) elicitation methods are shaped more by system characteristics and community-specific practices than theoretical commitments, particularly for predictive and explainable AI systems and autonomous or driver-assist vehicles. We derive 9 practical guidelines to support more deliberate and reflective methods for eliciting mental models of AI systems. In doing so, we aim to reestablish continuity between the cognitive theory of mental models and their empirical use in HCI, improving the transparency and comparability of research surrounding the concept.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; **HCI theory, concepts and models**; • **General and reference** → **Surveys and overviews**.

Keywords

Mental model, human-centered AI, systematic literature review, elicitation methods, guidelines

ACM Reference Format:
Téo Sanchez, Oleksandra Vereschak, and Ophelia Deroy. 2026. Mental Models in Human-AI Interaction: Systematic Review of Empirical Methodologies and Guidelines. In *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3742413.3789223

IUI '26, Paphos, Cyprus
© 2026 Copyright held by the owner/author(s).
This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus*, https://doi.org/10.1145/3742413.3789223.

1 Introduction

The concept of mental model has been developed to describe how people represent and reason about the world. Theorized in the 1980s in cognitive psychology [59], the concept was soon adopted in human-computer interaction (HCI) to approach users' reasoning about computing devices [101]. Mental models are internal, condensed, and personally constructed representations of external phenomena or systems [119]. A mental model is more than a collection of facts; it is an internal proxy of a system, whose simulation scaffolds reasoning and enables new inferences to be made. Importantly, mental models may be inaccurate and parsimonious [101], because they are constructed and updated from people's own experiences. Eliciting users' mental model of computing systems has been instrumental in HCI research, because it can explain *how* and *why* people use, rely on, trust, or reject these systems. For instance, believing that voice assistants eavesdrop on private conversations for targeted advertising [81, 128] can shape users' interaction and trust toward these systems, even when such a belief is incomplete or inaccurate. Ultimately, eliciting users' mental models of computing systems may inform their design and make them more intuitive and discoverable. Yet, challenges have emerged as the concept migrated from a unified theory of thoughts in cognitive psychology [58] to a pragmatic construct in HCI research. Conceptually, researchers have proposed multiple definitions [119, 120] and theoretical refinements [8, 96, 116], often tailored to the needs of their disciplines. This may explain why empirical studies in HCI were found to invoke mental models loosely and uncritically [54, 112, 141], as a catch-all construct with little distinction as to users' general knowledge.

Beyond this conceptual noise, we argue that epistemological questions about **how mental models are elicited** are at least as important as ontological debates about **what they are**. Elicitation is the process of capturing a person's mental model in a form that can be examined by others [6]. It is an interpretative process of perspective-taking, assumed to be self-reflective (i.e., elicitation may change the mental model being elicited) and shaped by layers of biases [117], as illustrated in Figure 1. Elicitation methods are less contested than the concept of mental model, but present their own challenges and knowledge gaps. First, the repertoire of elicitation methods is large [20, 63, 79], which makes it difficult to determine which method to apply and why. Second, elicitation is not a neutral act of data collection, but embeds assumptions about what aspects of user cognition are made visible or remain hidden, and what is considered a mental model in the first place. Third, research

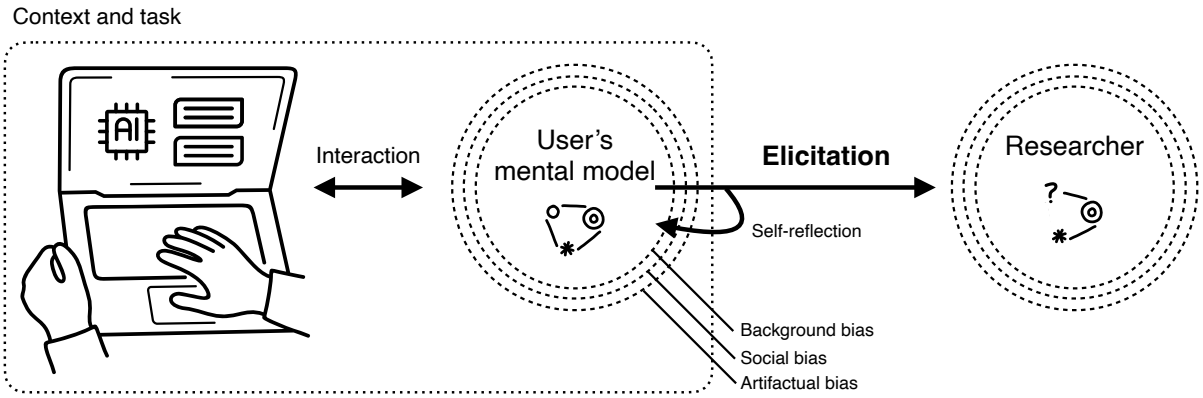


Figure 1: Mental model elicitation is an interpretive and self-reflective process, shaped by three layers of bias from both the participant and researcher, that shape which information is noticed, ignored, or assumed [117]: background biases (prior knowledge and expectations), social biases (influences during participant–researcher interaction), and artifactual biases (introduced by the elicitation substrate, such as interviews or diagrams). Adapted from Revell and Stanton [117].

on mental models has historically focused on interactive systems producing deterministic responses through direct manipulation of digital objects. By contrast, modern AI systems have distinctive capabilities (e.g., language and image understanding), forms (e.g., from anthropomorphized avatars to integrated in graphical user interfaces), and levels of autonomy [38] (e.g., AI agents). These properties may introduce new obstacles to human understanding and influence how researchers approach mental model elicitation.

Together, these challenges and the rapid adoption of consumer-facing AI services since 2022 highlight the need to reexamine how mental models of AI systems are elicited. Prior work listed elicitation methods [20, 63, 79] without mapping their actual use in empirical research. Only Hu and Twidale [54] conducted a scoping review of mental model research in HCI and looked at how studies define, elicit, and represent mental models across a wide range of interactive systems. None of these studies explores mental models in the context of human-AI interaction, nor examines the relation between elicitation, their epistemic assumptions (i.e., what kind of mental model is assumed), definitions, and the type of system studied. Understanding these relations is important because it could surface whether AI systems complicate the methodological praxis associated with mental models, and, in turn, support researchers in making more reflective and transparent methodological choices that support comparable and cumulative progress. To address this gap, our work investigates three research questions:

RQ1. How do researchers elicit users' mental models of AI-enabled systems, and what epistemic assumptions and trade-offs do these methods entail? **RQ2.** How do elicitation methods relate to mental model definitions and the type of AI systems investigated? **RQ3.** In what ways do the properties of AI systems complicate or reshape how researchers approach and elicit users' mental models?

To answer these questions, we conducted a systematic literature review of empirical studies that investigate users' mental models of AI-enabled systems. We retrieved 88 empirical papers and extracted structured information about the AI system under investigation, participants involved, the definitions of mental models, and, most importantly, the epistemic assumptions and elicitation methods employed, including methodological limitations reported by the researchers themselves.

Contributions. Our work makes four contributions. **(1)** We clarify how the mental model construct has been conceptualized across cognitive psychology and HCI, contrasting their underlying assumptions and theoretical framings (Section 2), and propose a clear typology of elicitation methods (Section 2.3). **(2)** Our systematic review extends previous findings [54], by showing that elicitation methods are shaped more by system characteristics and community-specific practices than theoretical commitments, particularly for predictive and explainable AI systems and autonomous or driver-assist vehicles. **(3)** We distill practical guidelines to support more deliberate and reflective methods for eliciting mental models of AI systems (Section 5.2). Along with its supplementary material, our work can serve as a reference point for scholars to identify, reflect, and improve upon existing methods, ultimately supporting transparency and comparability of research about mental models of AI systems.

2 Background and related work

Prior work described the appeal from the mental model concept in HCI research [112, 141], but also its loose and inconsistent use [54]. Although Johnson-Laird's foundational book [58] in cognitive psychology is frequently cited [54], its conceptual components and properties are not always reminded. Subsection 2.1 summarizes Johnson-Laird's theory and outlines the distinctive properties of mental models as framed in cognitive psychology. Subsection 2.2 then examines the concept's adoption and framing within HCI research. Subsection 2.3 introduces a typology of elicitation methods to structure data collection of our literature review. Finally, Subsection 2.4 summarizes the research gap that motivates our work.

2.1 Mental models as a unified theory of reasoning in cognitive psychology

In *"Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness,"* Johnson-Laird [58] introduced a unified theory of reasoning that rejects what he called the "doctrine of mental logic", i.e., the idea that people reason by applying formal syntactic rules to reason about the world, akin to propositional calculus. Under the logicist view, a conditional statement like "If P then Q" would be processed independently of the content of P and Q once the deduction rule is acquired. Simple experiments such as the Wason selection task [89, 139] have challenged the logicist view by revealing systematic reasoning fallacies called *"illusory inferences"* [61]. The mental model theory aims to explain several empirical observations, mainly from reasoning tasks, and relies on three components: propositional representations, the procedural semantics, and schemata.

Propositional representations. Expressible propositions are mentally encoded in so-called *propositional representations*—basic units of meaning and building blocks of mental models. Propositional representations do not support reasoning beyond the information they contain and act as static pieces of knowledge. Only when integrated into a mental model can they enable reasoning beyond what is explicitly stated [58].

Procedural semantics. The construction and revision of mental models is guided by procedural semantics—a set of cognitive operations that assemble, revise, and integrate propositional representations into mental models. These procedures determine how new propositional representations are integrated, inconsistencies are resolved, and how new propositions are inferred during reasoning [58].

Schemata. Mental models are assumed to be influenced by preexisting and schematic knowledge structures—referred to as *schemata*[1] [121], and *scripts* in social contexts [125]. Schemata are stored in long-term memory and encode prototypical rather than episodic experiences. They can be understood as templates for the creation of mental models.

Mental models. Johnson-Laird [58] proposes that mental models are *structure-preserving*: their internal organization mirrors the structure of the real or imagined system they represent, rather than the surface-level structure of the linguistic cue. Preserving structure allows mental models to be *cognitively efficient*, facilitating the integration and retention of information. Then, mental models are *inferential*, enabling individuals to go beyond the information encoded and draw new conclusions by "running" the model, much like a proxy of reality. Lastly, they are *context-sensitive*, emerging from users' prior and situated knowledge rather than from formal logical reasoning. Consider a typical example inspired from Mani and Johnson-Laird [88] and describing a table setting: *"The glass is above the plate. The cutlery is next to the plate. The knife is to the right of the fork."* Rather than remembering this description verbatim, people tend to construct a spatial representation of the described layout [88]. This mental model preserves the spatial relationships (structure-preserving), is easier to remember and update than the verbatim textual description (cognitively efficient) [88], and enables one to guess that *"the fork is left to the knife"* (inferential). Readers with Western schemata of table-setting may construct a model in which the fork and knife are placed on either side of the plate, although this aspect remains undetermined in the description above (context-sensitive). In sum, cognitive psychology theorized mental models from controlled reasoning experiments, and as Johnson-Laird put it, *"reasoning is more a simulation of the world fleshed out with all our relevant knowledge than a formal manipulation of the logical skeletons of sentences"* [60]. Acknowledging mental models' distinctive properties—*structure-preserving*, *cognitively efficient*, *inferential*, and *context-sensitive*— can help HCI practitioners identify discrepancies in the way mental models are operationalized in empirical research.

2.2 Mental models in human-computer interaction

Aside from a few theoretical extensions of Johnson-Laird's work in human–machine contexts [8, 96, 116], most of the mental model's uptake in HCI has been pragmatic, shaped by design and usability concerns. First, mental models if often reified as laypeople's beliefs about computing systems, overlapping with the concept of *folk theories*. The latter align more with social psychology rather than cognitive psychology traditions and emphasize shared and culturally transmitted beliefs, without necessarily assuming the structure-preserving and inferential features described in the previous section. Second, eliciting users' erroneous reasoning about computing systems has been seen as a way to inform the design and instruction materials of these systems. However, the passage from mental model elicitation to design improvements has been criticized to be more often assumed rather than demonstrated [141]. Norman [100] notoriously argued that it's the role of the designer to shape the system's image to guide users toward constructing an accurate mental model. Third, researchers commonly distinguish between two kinds of users' mental models when applied to machines, though they appear under diverse terminologies[2] [72, 76, 146]: **functional mental models** capture a system's behavior at a surface-level and how to operate a system, and **structural mental models** capture how a system works internally. This distinction can be traced back to Young [146], who similarly introduced *mappings* and *surrogates*. Mappings (functional mental model) describe the relationship between user actions and device response, while surrogates (structural mental model) serve as a notational or physical analogue of the device's internal mechanism[3]. Table 1 compares the different and confusing terminologies for these two recurring categories of mental model. The functional–structural distinction has occasionally been used as an experimental manipulation [71, 76, 146]—training participants to develop one type of mental model or the other and then measuring performance. When these lenses instead underlie

[1]The notion of schemata has also been subject to theoretical divergences [4]. Please refer to Jones et al. [63] for a more comprehensive comparison in the context of mental model.

[2]In the rest of the paper, we align with the terminology from Kulesza et al. [76], chosen for its legibility and simplicity.

[3]This distinction also echoes Johnson-Laird's typology of mental models. Functional models align with his notion of physical-temporal models, which represent observable events as ordered sequences (e.g., "swiping down refreshes the page"). Structural models correspond to conceptual-relational models, which capture abstract structures and relationships, such as data flows or system architectures.

Field	Source	Type of mental model: **Functional** (operations)	**Structural** (inner working)
HCI	Young, 1983 [146]	Mapping	Surrogate
	Kieras and Bovair, 1984 [72]	User's mental model	Device model
	Bayman and Mayer, 1984 [10]	Mental model	Mental model
	Kulesza et al., 2012 [76]	Functional	Structural
Cognitive psychology	Johnson-Laird, 1983 [58]	Physical-temporal	Conceptual-relational

Table 1: Terminologies used to distinguish functional and structural mental models across the literature. The last row maps this distinction to Johnson-Laird's typology.

the elicitation itself, however, they frame different facets of reasoning and can make results across studies difficult to compare. This incommensurability also motivates our systematic mapping of the epistemic assumptions underlying empirical elicitation methods.

Lastly, an important pragmatic consideration emphasized by Norman [101] is that users' mental models are often incomplete, imprecise, and parsimonious, and claims that users are often aware of their own uncertainties and reasoning limitations about the target system. This parsimony reflects the *cognitive efficiency* of mental models described in the previous subsection, i.e., mental models only represent the information needed to reason and act within context[4]. In Norman's words, scientists *"must develop appropriate experimental methods and discard our hopes of finding neat, elegant mental models, but instead learn to understand the messy, sloppy, incomplete, and indistinct structures that people actually have."*

2.3 A typology of mental model elicitation methods

Andrews et al. [6] defines mental model elicitation as the process of capturing a person's mental model in a form that can be examined by others. We adhere to this simple definition, but distinguish elicitation from two other methodological interventions often conflated [29, 54, 63, 80]: post-hoc analysis and behavioral measurements. Post-hoc analysis transforms elicited data into more exploitable forms, often for quantification and comparison; and performance-oriented behavioral measurements[5] usually captures the consequence (e.g., performance) of mental models rather than producing an explicit and inspectable representation of a mental model's structure or content. We propose a typology of four elicitation groups described below:

Researcher-led verbalizations involve participants articulating their mental models verbally, in response to prompts prepared by the researcher, typically during interviews or free text questions in surveys. These prompts may guide participants through different modes of reasoning, such as recalling specific events, providing general explanations, using metaphors, and commenting on hypothetical or counterfactual situations [68, 111]. We distinguish this group from participant-led verbalizations as these techniques may encourage forms of reasoning that participants might not spontaneously use otherwise (e.g., counterfactual).

Participant-led verbalizations involve participants expressing their understanding with little researcher guidance. These techniques may capture reasoning as it unfolds—either during participants' interaction with a system (e.g., think-aloud protocols [65]) or retrospectively (e.g., commentary or teach-back [36]). They are especially effective for surfacing moment-to-moment decisions and metacognitive reflections. As Ericsson and Simon [33] notes, think-aloud protocols aim to externalize inner speech and often yield unstructured or fragmented data, whereas teach-back and interviews involve social speech and usually produce more coherent verbalization.

Artifact-based elicitation methods involve participants manipulating meaning-laden items—typically represented as cards—that stand for concepts, system components, or features. Cards echo Johnson-Laird's propositional representations, and their sorting, grouping, or mapping externalize the causal and structural nature of mental models. More creative and open forms of artifact-based elicitation may invite participants to create visual representations through drawing, mock-ups, storyboards, etc. This group of elicitation is often engaging, produce persistent and easily reviewable artifacts, and can be readily combined with—or even facilitate—verbalization (e.g., think-aloud).

Structured judgments and ratings use predefined formats—such as scales, multi-answer questions, or comparison grids—to elicit participants' judgments about system components or features, typically identified in advance. These methods yield structured and comparable data but usually focus on discrete attributes evaluated in isolation, making them less suitable for capturing the relational and inferential nature of mental models.

2.4 Research gap

To our knowledge, Hu and Twidale [54] provides the only literature review mapping how empirical HCI studies define and elicit users' mental models of computing systems. While valuable, their analysis has limited relevance for our research questions for three reasons. First, it spans a wide range of systems (e.g., gesture control, cybersecurity, virtual reality) in which AI represents only a small portion of the corpus. This limitation is amplified by the fact that their review predates the mainstream adoption of generative AI

[4]This idea also echoes Herbert Simon's notion of *bounded rationality* [129]

[5]Distinguishing behavioral measures from elicitation may seem counterintuitive, as cognitive psychology [88, 139] generally treats behavior as evidence of underlying cognitive functions. However, the classic reasoning tasks [62, 88, 89] supporting the mental model theory typically test the presence of certain types of mental models rather than revealing their content, what elicitation methods in HCI aim to do.

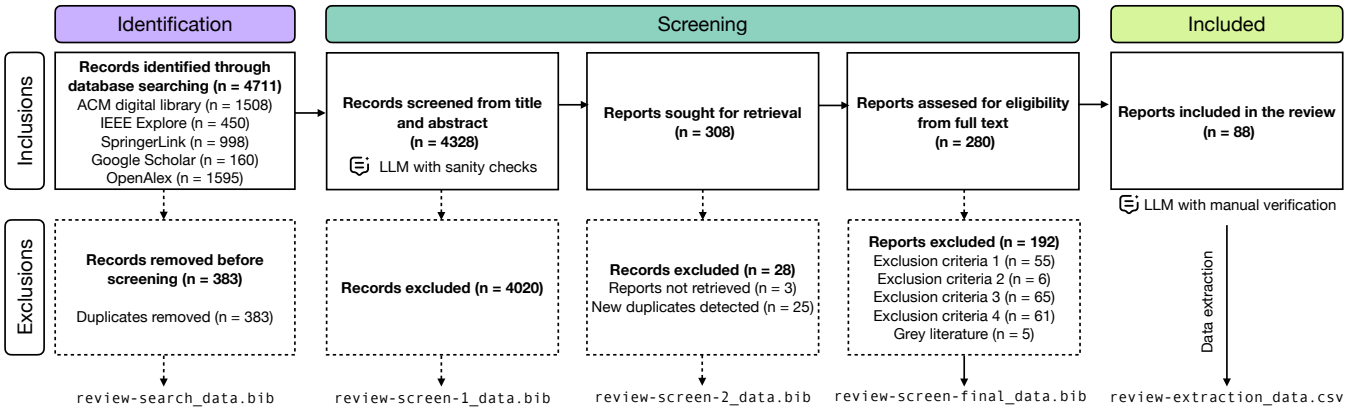

Identification
Screening
Included
Inclusions
Exclusions
Records identified through database searching (n = 4711)
ACM digital library (n = 1508)
IEEE Explore (n = 450)
SpringerLink (n = 998)
Google Scholar (n = 160)
OpenAlex (n = 1595)
Records screened from title and abstract (n = 4328)
LLM with sanity checks
Reports sought for retrieval (n = 308)
Reports assesed for eligibility from full text (n = 280)
Reports included in the review (n = 88)
LLM with manual verification
Records removed before screening (n = 383)
Duplicates removed (n = 383)
Records excluded (n = 4020)
Records excluded (n = 28)
Reports not retrieved (n = 3)
New duplicates detected (n = 25)
Reports excluded (n = 192)
Exclusion criteria 1 (n = 55)
Exclusion criteria 2 (n = 6)
Exclusion criteria 3 (n = 65)
Exclusion criteria 4 (n = 61)
Grey literature (n = 5)
Data extraction
review-search_data.bib
review-screen-1_data.bib
review-screen-2_data.bib
review-screen-final_data.bib
review-extraction_data.csv

Figure 2: Flow diagram summarizing the systematic review's phases and following the PRISMA 2020 reporting guidelines [108]. The diagram reports the number of records identified, included, and excluded, along with exclusion reasons. Steps assisted by large language models are annotated (see corresponding sections for details), as well as the names of the resulting files, available in the supplementary material.

systems in 2022. Second, it does not examine the methods' epistemic assumptions, i.e., what kind of mental model is assumed or made visible through a given method. Third, it maps methods and definitions but does not explore whether these relate to each other, and to the type of system being studied. Revealing these patterns or discrepancies in human-AI interaction can have two important consequences. First, it can surface whether and how AI systems complicate or reshape the methodological praxis associated with mental models. Second, it can help HCI researchers reflect on their own practices and assumptions when using the concept, ultimately leading to more transparent and comparable research findings.

3 Methods

The previous sections outlined the conceptual foundations of mental models in cognitive psychology and HCI, introduced a typology of elicitation methods, and identified the research gap. Building on this background, our goal is to examine how mental models are studied within human-centered AI research. To this end, we conducted a systematic review of empirical studies investigating users' mental models of AI systems. This section describes our search strategy, selection process, data extraction, and synthesis procedures. The review adheres to PRISMA reporting guidelines [108] where applicable and was pre-registered on https://osf.io/54brg/overview[6].

3.1 Search strategy

Our keyword search strategy emulates Vereschak et al. [136], who reviewed methodologies on trust in AI-assisted decision-making. We, however, extended the search scope to five digital libraries: ACM Digital Library, IEEE Xplore, OpenAlex, Springer Link, and Google Scholar. Search queries were tailored to each platform's capabilities and detailed in Table 2 in Appendix A. In general, the search strings combine keywords related to AI and to empirical studies of human mental models, applied to the full text and abstract, where supported. The search was conducted on May 21st, 2025, with no publication date restrictions, and yielded 4,328 records after de-duplication[7].

3.2 Selection process

The selection process—including records' identification, screening, and final inclusion—is summarized in Figure 2. Screening proceeded following the two standard phases. First, records were screened by title and abstract, prioritizing sensitivity, ensuring that all potentially relevant papers were captured, even at the expense of retrieving some irrelevant results. Second, included reports were retrieved (when possible) for full-text screening. Full-text screening prioritizes specificity, ensuring that only relevant papers are captured and thus determines the final corpus. In both screening phases, documents were assessed against the four inclusion criteria, listed below in increasing specificity to our research goals:

C1 Format: The publication must be a full-length research paper to ensure sufficient methodological detail. Posters, extended abstracts, late-breaking works, workshop papers, and publications in languages we can not read should be excluded. **C2 Empirical study with human participants:** The publication must include an empirical study involving human participants. Theoretical papers, design explorations, and simulation-based studies without direct human involvement should be excluded. **C3 AI-enabled system:** The research paper must focus on AI-enabled systems, broadly defined to include interfaces, robots, virtual agents, or autonomous agents, as in prior reviews [45, 136]. Articles that do not consider AI-enabled systems should be excluded. **C4 Mental model focus:** The research paper must examine or elicit humans' mental models

[6]Deviations from the pre-registration include: (1) a greater use of LLMs to assist data extraction on open-access or author-version papers only, with all outputs manually verified; (2) the use of Zotero instead of Taguette for simplified collaborative annotation; (3) the exact relational analyses for multi-categorical variables were not anticipated; and (4) variable definitions and possible values were iteratively refined during extraction.

[7]De-duplication was performed with the systematic review management platform Rayyan [107], which includes a duplicate's automated detection functionality with manual confirmation.

of an AI-enabled system. Papers that do not explicitly employ mental models as a methodological emphasis or object of investigation should be rejected.

We employed large language models (LLMs) to assist with title and abstract screening. Recent studies indicate that, under appropriate conditions, LLMs can achieve comparable [57, 69] or even superior [19, 64, 123] performance to human screeners. First, unlike certain clinical reviews, our research question can tolerate a low number of false negatives, i.e., papers excluded that should have been included. Second, we drew on the best practices from the emerging literature of LLM-assisted screening: (1) prompts were optimized and records anonymized to minimize bias [19]; (2) screening used an ensemble of three LLMs with majority voting; and (3) human sanity checks were conducted to assess reliability. More specifically, we refined the Abstract ScreenPrompt template proposed by Cao et al. [19] and screened records using three LLMs—qwen3:14b, gemma3:27b-it-qat, and phi4:14b— hosted locally and run with Ollama[8]. For each record, LLM's inclusion decisions were determined by majority vote across the three models. To assess the reliability of this procedure, we conducted a sanity check on a random sample of 200 records (100 included and 100 excluded by the LLM ensemble). The first author reviewed these records using Rayyan [107], and human judgments were treated as ground truth for this sample. Agreement between the LLM ensemble and the human reviewer reached 81% (163/200), with a precision of 0.76 and a recall of 0.85. The sanity check was intended to estimate agreement between LLM-assisted and human screening rather than to retroactively correct screening decisions across the full corpus. Nevertheless, the 37 disagreements within this sample were adjudicated by the last author, resulting in 28 rejections and 9 inclusions that were carried forward to subsequent screening stages. Full-text screening was then conducted manually by the first two authors in a single pass, with uncertainties resolved through deliberation. From the 4,328 articles initially identified, 88 met all inclusion criteria after screening.

3.3 Data extraction

We adopted a structured coding scheme to guide data extraction from the included papers. Although not directly addressing our research questions, some variables were included to support contextual analysis and assess the corpus's quality (e.g., metadata, participants' details, quality indicators). The extracted variables were grouped as follows:

V1 **Metadata**: title, DOI, authors, venue, and publication date; V2 **AI system characteristics**: based on a simplified version of the OECD classification of AI systems [105], including system description, application domain, type of AI system, system's task, level of user control [38], and system's image [100] (e.g., graphical user interface, speaker, avatar etc.); V3 **Participants' attributes**: sample size, recruitment method, country of recruitment, age range, domain of expertise, and profession; V4 **Mental model definitions**: definition used (verbatim), definition's key attributes (coded from the definition), and cited sources; V5 **Methods**: type of study, administration modality, type of mental model assumed[9], scope of the mental model (e.g., specific system or class of systems), description of the elicitation method, elicitation method (coded from the description), methodological source cited, type of of elicitation (from our typology in section 2.3), description of post-hoc analysis, whether participants were exposed to the system during the study, whether group comparisons or updates of the mental model were examined, author-reported limitations (verbatim), specific and general types of author-reported limitation (coded from author-reported limitations); V6 **Quality indicators**: use of reliability psychometrics, pre-registration, ethical approval, and availability, type, and location of supplementary materials. A complete description of each variable is provided in the supplementary material.

Data extraction was supported by large language models (LLMs), primarily to expedite the identification of relevant passages within the full text. For each variable, the LLM was requested to extract the values, justifications, and verbatim evidence from the text. Only open-access or author-version papers were processed using gpt-5-mini (low verbosity, medium reasoning effort), while copyrighted papers were analyzed manually. All extracted outputs were subsequently verified by the first two authors within the documents, with particular attention to the most interpretative or critical variables (e.g., assumed type of mental model, mental model elicitation methods). The final structured dataset, including all extracted variables, is provided as a supplementary spreadsheet.

3.4 Synthesis process

The synthesis aggregates and interprets extracted data across papers. We followed three main approaches.

Qualitative coding. Interpretative variables—such as the type of mental model assumed, elicitation methods, and reported limitations—were inductively analyzed during data extraction. New values were added when encountered and later merged or renamed to ensure coherence across the dataset. This iterative process balanced openness to emerging categories with consolidation of overlapping codes.

Descriptive analysis. Extracted data were first summarized within each variable group using descriptive statistics and visualizations (e.g., counts, proportions, and bar charts). To characterize the group V2 **AI system characteristics**, we applied k-modes clustering to three categorical variables—application domain, type of AI system, and system task. This method was chosen under the assumption that AI systems cannot be meaningfully described by technical properties alone, but must also account for their application context. The number of clusters was iteratively increased from $k = 5$, with manual inspection of clusters' distribution and coherence at each step, until $k = 7$. Two similar clusters were then merged, resulting in six distinct groups of AI systems, described in Section 4.2.

Relational analysis. To examine relationships between multi-categorical variables (e.g., the link between elicitation methods used and type of AI systems studied), we constructed co-occurrence matrices in which rows and columns represented all possible categories of two variables, and each cell recorded the number of

[8] https://ollama.com/

[9] Coding of the type of mental model assumed reflects the mental model presupposed by a paper's elicitation methods, rather than participants' reported responses.

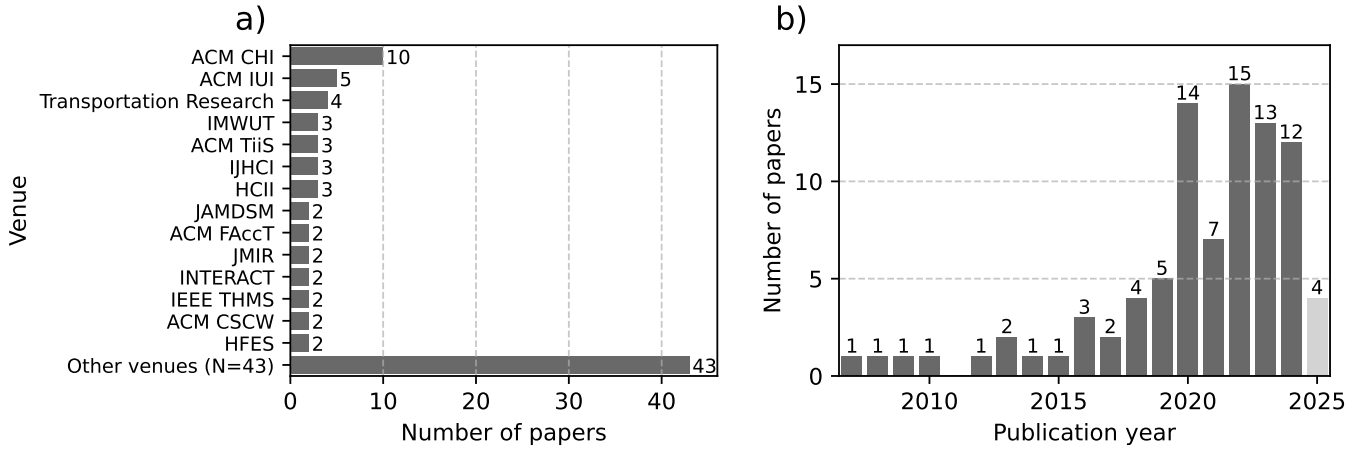


Figure 3: Number of included studies by (a) publication venue and (b) publication year. The 2025 bar is shown in a different color as data were collected mid-year and are not yet representative.

papers in which a given pair of values co-occurred. We then applied a Chi-square test of independence (χ^2) to assess whether observed frequencies deviated from the null hypothesis of independence. If $p > 0.05$, variables were considered unrelated and not analyzed further. Otherwise, we conducted pairwise Fisher's exact tests between all category pairs, applying a Benjamini–Hochberg correction for multiple comparisons. We report results from this relational analysis either in text or through a matrix showing significant co-occurrence (see Figure 6), only highlighting associations with an adjusted threshold of $p < 0.1$ to reflect the exploratory nature of our analysis.

Authors' background: The first two authors work at the intersection of HCI and human-centered AI, with expertise in interactive ML and trust in AI-assisted decision making. The last author specializes in philosophy of mind with extensive background in cognitive psychology. All authors hold doctoral degrees.

4 Results

This section presents the results of the systematic review. We first describe the composition of the corpus (Section 4.1) and the types of AI systems examined (Section 4.2). We then analyze how studies define the notion of *mental model*, situating these definitions within the dual psychology–HCI legacy of the concept and examining whether definitional choices relate to methodological ones (Section 4.3). Section 4.4 details how mental models are elicited across the corpus, including the scope of the system considered, whether system-specific or class-level (Section 4.4.1), the epistemic assumptions each method presupposes (functional, structural, or analogical/anthropomorphic; Section 4.4.2), the elicitation techniques employed (Section 4.4.3), and the limitations reported by authors (Section 4.4.4). We report on co-occurrence relationships between variables following the analysis pipeline described in Section 3.4. Finally, Section 4.5 summarizes indicators of methodological quality and transparency. Important findings are highlighted with F markers.

4.1 Corpus overview

Figure 3 summarizes the distribution of the 88 papers included in our review, by publication venue and year. HCI venues dominate the sample with 52%[10] of the studies published, followed by AI (11%) and automotive-oriented (9%) venues. The remaining studies appear in domains such as robotics, education, health, and psychology. This distribution is not surprising considering our inclusion criteria (AI systems and human participants), and confirms earlier observations about the sustained use of the mental model concept in HCI [54]. The temporal distribution shows an increase in publications from 2020 onward. This increase likely reflects the intensifying research efforts and funding on AI-related topics [39], which might certainly be visible in publication trends within HCI venues. Research about mental models of AI systems might also be motivated by the widespread adoption of consumer-facing generative AI technologies since 2022.

4.2 Type of AI systems studied

The corpus examines a diverse range of AI-enabled systems that cannot be easily characterized along a single dimension. To structure this diversity, we drew on a simplified version of the OECD classification of AI systems [105] and extracted six variables—mostly multi-categorical—as described in Section 3.3. We capture higher-level structure about AI systems by applying a k-modes clustering algorithm ($k = 7$) on three variables: application domain, system type, and system task. K-modes is an optimization algorithm using a simple matching dissimilarity measure to partition n observations (the papers) into k clusters (the system types) such that each paper belongs to the cluster with the smallest dissimilarity. After manual inspection and merging two similar clusters together, we obtained six clusters. Clusters' distribution is shown in Figure 4, and their characteristics are described below.

[10]Percentages are rounded up to the nearest whole number

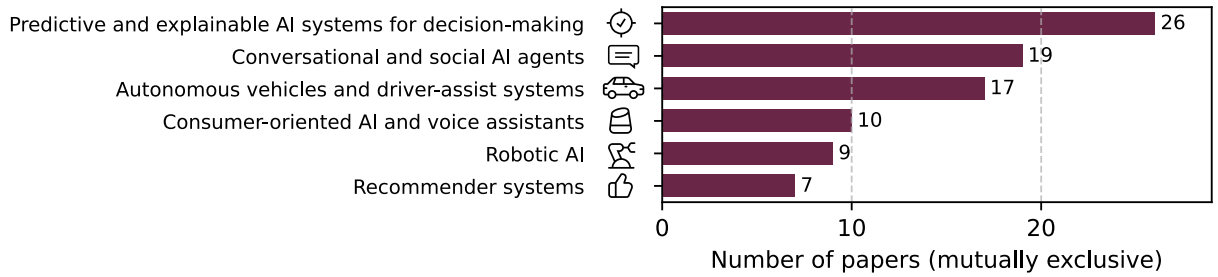


Figure 4: Distribution of the AI systems studied in the corpus. The six groups were identified using k-modes clustering based on each system's type, task, and application domain. Counts are mutually exclusive as each paper is uniquely assigned to one cluster.

- **Predictive and explainable AI systems for decision-making**[11] refers to predictive and decision-support systems (54%) performing classification, recognition, or detection (65%), decision support and diagnosis (23%), and prediction or forecasting (15%), mostly through a graphical user interface (92%). Papers span diverse application domains, including healthcare (27%), consumer technology (23%), entertainment and media (15%), and workplace, teamwork, and productivity (12%). At least nine papers can be situated within the research area of explainable AI [5, 27, 44, 85, 93, 102, 122, 130, 140].
- **Conversational and social AI agents**[12] encompass textual conversational agents (74%), mainly performing text and synthetic media generation (63%) and information retrieval (37%) across diverse application domains. Systems mainly take the form of graphical user interfaces (53%) or virtual agents or avatars (32%).
- **Autonomous vehicles and driver-assist systems**[13] form the most homogeneous cluster, exclusively addressing transportation and mobility with autonomous vehicle or driver-assist systems (ADAS). Importantly, this cluster also reflects a research community operating apart from generalist HCI or AI publishing venues.
- **Consumer-oriented AI and voice assistants**[14] mainly account for voice assistant (90%) and context-aware and activity recognition system (10%) and apply to consumer technology (100%). Main tasks include general assistance (70%) and information retrieval (30%). Systems mainly take the form of speakers, IoT devices (80%), or smartphone applications (10%). Although voice assistants are a type of conversational agent, this cluster exclusively includes voice-based and consumer-oriented systems (e.g., Alexa), whereas the conversational and social AI agents cluster mainly includes text-based agents and applies to various domains.
- **Robotic AI**[15] describe AI-embeded robots (89%), either physical or virtual, that perform a wide range of tasks, including exploration and search (33%), classification, recognition, or detection (22%), planning and coordination (22%), social influence and emotional support (22%), and skill acquisition or learning (11%).
- **Recommender systems**[16] account for recommender systems (86%) performing recommendation or personalization (100%) and applying to entertainment and media (100%). Systems mainly take the form of graphical user interface (86%) or virtual agents and avatars (14%).

These six clusters summarize the diversity of AI systems represented in the corpus and serve as a basis for analyzing how different system types relate to researchers' definitions, assumptions, and elicitation of mental models.

4.3 Mental model definitions

Our analysis shows that **F1** **mental model definitions, when provided, are grounded in system-oriented interpretations and rarely highlight all the distinctive features of the mental model theory**, aligning with earlier observations [54]. We extracted verbatim definitions of *mental model* from the papers and identified their sources and key attributes. One third of the papers (34%) do not provide any definition—echoing prior findings by Hu and Twidale [54]. The five most cited definitional sources are Norman's *Some Observations on Mental Models* [101] (11 citations) and *The Psychology of Everyday Things* [100] (10 citations), Johnson-Laird's foundational book *Mental Models* [58] (9 citations), and two HCI-oriented syntheses by Rouse and Morris [120] and Carroll and Olson [22] (6 each). The distribution of definitional sources shows a long tail, with 42 definition sources cited only once across the corpus. This finding highlights the predominance of applied, system-oriented interpretations of mental models, with comparatively fewer direct references to cognitive psychology. When provided, definitions are diverse and emphasize certain definitional attributes we identified and listed below by decreasing prevalence:

- **Explanatory and causal** (45%)[17]: mental models are described as supporting explanation and reason about cause–and–effect relationships, e.g., *"mental models have been defined as a representation of the typical causal interconnections involving actions and environmental factors that influence a system's functioning"* [56];

[11] Predictive and explainable AI systems for decision-making: [5, 9, 18, 27, 35, 41, 44, 49–52, 70, 85, 93–95, 102, 103, 115, 122, 126, 130, 132, 134, 137, 138]
[12] Conversational and social AI agents: [2, 14, 15, 23–25, 31, 32, 40, 43, 47, 66, 67, 73, 91, 106, 140, 145, 148]
[13] Autonomous vehicles and drive-assist systems: [7, 11, 12, 21, 26, 37, 55, 56, 74, 84, 97, 98, 104, 114, 127, 147]
[14] Consumer-oriented AI and voice assistants: [1, 28, 34, 53, 82, 83, 87, 109, 133, 142]
[15] Robotic AI: [75, 86, 92, 110, 113, 124, 131, 143, 144]
[16] Recommender systems: [3, 48, 76–78, 90, 99]
[17] Percentage calculated over the subset of papers providing a definition.

- **Domain-specific references** (43%): definitions tailored to particular application contexts, e.g., *"a mental model, in this context, refers to a driver 's knowledge and understanding of the scope, functionality, and rationale behind the behavior of an automated driving system."* [147];
- **Internal/cognitive** (43%): mental models are defined as internal cognitive representations, rather than observable behaviors, e.g., *"a mental model is a cognitive representation that a user has about a complex model"* [52]:
- **Predictive/inferential** (41%): definitions emphasizing mental models' role in supporting prediction and inference, e.g., *"mental models are simplified representations of the world that people use to process new information and make predictions"* [67];
- **Experience-based and dynamic** (23%): definitions describing mental models as evolving with user experience and interaction, e.g., *"mental models are internal representations that people build based on their experiences in the real world"* [5];
- **Representational scope** (17%): define mental models in terms of what they can represent, e.g., *"mental models, in cognitive theory, provide one view on how humans reason either functionally (understanding what the robot does) or structurally (understanding how it works)"* [40];
- **Bounded and incomplete** (12%): definitions acknowledging mental models' partial and approximate nature, e.g., *"mental models are naturally evolving models that need to be functional but need not be technically accurate."* [110]

Beyond these descriptive patterns, the relational analysis described in Section 3.4 demonstrates that **F2** **definition choices are independent from methodological ones and from the type of AI system considered.** According to the Chi-squared test of independence, no significant relationships were found ($p > 0.05$) between these definitional attributes and other key variables outlined in other subsections, such as the type of AI system studied, the type of mental model assumed, or the type of elicitation method employed. This suggests that definitional choices are not tied to particular applications or methodological approaches, but may instead reflect heterogeneous appropriations of the concept. Some of these definitional attributes echo the four distinctive features of mental models highlighted in Section 2.1 but are rarely used altogether within a single paper. Without restating the broader cognitive psychology account at every new HCI study, attending to relevant attributes when defining mental models—such as internality, explanatory and inferential power, experience-based, and boundedness—can already sustain continuity between HCI and cognitive psychology.

4.4 Mental model's elicitation methods and their assumptions

4.4.1 Scope. Our analysis shows that **F3** **the scope of the elicitation, i.e., targeting a specific system or a class of systems, reflects a divide between mental models used for system evaluation and for lay understanding of AI.** Elicitation scope concerns what participants are expected to form a mental model *about*. We coded whether researchers elicit the mental model of a **specific system** (e.g., Alexa Echo Show [34]), a **class of systems** (e.g., voice assistants [109]), or a more **imagined or speculative system** [110]. We found that most studies elicit participants' mental models of a specific system (70%), one to which participants are predominantly exposed during the study (80%). According to the relational analysis pipeline described in section 3.4, co-occurrences between the scope and the type AI systems studied deviated from the null hypothesis of independence ($p_v < 0.01$), with mental models of specific systems predominantly elicited for predictive and explainable AI systems for decision-making ($p_v < 0.1$). Conversely, general classes of systems were predominantly elicited for consumer-oriented AI and voice assistants ($p_v < 0.1$). This distinction shows that the same concept of mental model may be operationalized differently, as an evaluative tool (e.g., assessing the interpretability of a new explainable AI technique [93]) or as a tool to understand shared perception of AI in everyday contexts [87] (e.g., understanding voice assistant's perceived roles in domestic life [87]), blurring the lines between *mental model* and *folk theory* and weakening the concept's theoretical strength.

4.4.2 Type of mental model assumed (epistemic assumption). We found that **F4** **epistemic assumptions exceed the classic functional and structural lens (how it behaves versus inner working) with analogical and anthropomorphic framings**. We use the term epistemic assumption to describe the type of mental model the study presupposes through its methodological approach, i.e., what kind of mental model researchers seek to elicit, irrespective of the mental model definition used. In very rare cases, these assumptions are stated explicitly, as for Garcia et al. [40]: *"the mental model questions aim to capture two different dimensions of the user's mental model: structurally, so how AUVs work (MM-Q1) and functionally so what the AUVs were doing (MM-Q2). "*. More often, we interpreted these assumptions from the study's procedures and instruments. Our analysis identified a third epistemic lens extending beyond the common functional-structural distinction described in section 2.2: *analogical and anthropomorphic mental models*. We treated anthropomorphism as a particular kind of analogy in which the AI system is framed as a social entity with a role and intentions (e.g., a collaborator, an assistant), or further endowed with human-like traits (e.g., warmth, empathy, and morality). Figure 5 shows the distribution of these epistemic assumptions across papers. Functional assumptions are most common (85%), followed by structural (43%),

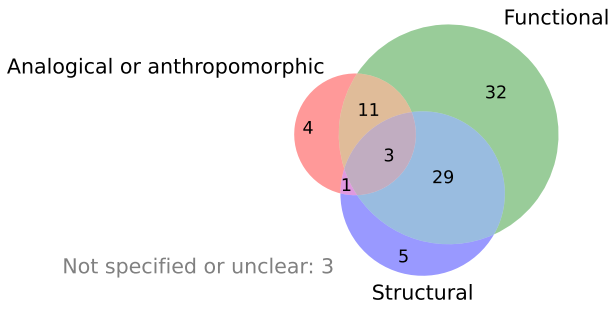


Figure 5: Types of mental model assumed by their elicitation methods across papers. Overlaps indicate papers targeting several types.

and analogical or anthropomorphic assumptions (22%). Functional assumptions are often combined with structural ones in 36% of the corpus. Conversely, structural assumptions are almost never combined with analogical or anthropomorphic ones.

Beyond these descriptive patterns, the relational analysis demonstrates that **F5** **functional, structural, and analogical mental models assumptions relate to types of AI systems studied, suggesting community-specific practices.** According to the Chi-squared test of independence, co-occurrences between epistemic assumptions and the types of AI system studied deviated from the null hypothesis of independence ($p_v < 0.01$). Consequently, Figure 6 reports the significant co-occurrences ($p_v < 0.1$) according to pairwise Fisher's exact tests (Benjamini-Hochberg corrected). A paper-level visualization of these co-occurences is provided in

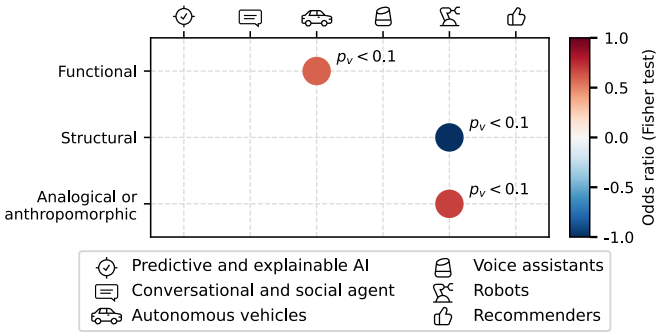


Figure 6: Co-occurrence analysis between the type of mental model assumed and the type of AI system studied (clusters). Only associations showing $p_v < 0.1$ on Fisher's exact test are represented. Circles' color encodes the odds ratio: red indicates that the row-column co-occurrence is over-represented in the corpus, while blue indicates under-representation.

Appendix B, showing how epistemic assumptions within papers distributes across the types of AI system considered. The analysis reveals that (1) studies about autonomous and driving-assisting vehicles predominantly assume functional mental models, and (2) studies about AI-enabled robots predominantly assume analogical and anthropomorphic mental models and are less likely to assume structural ones. While some associations are expected (humanoid robots invite anthropomorphic elicitations), others may reflect research-community boundaries, where particular AI applications converge on shared methodological assumptions. As AI technologies increasingly blur the line between tools and agents, clarifying the kind of mental model being elicited—functional, structural, analogical, anthropomorphic, or several at once—becomes essential to developing incremental and comparable results.

4.4.3 Mental model's elicitation methods. We found that **F6** **most paper triangulates elicitation methods, mainly drawing from researcher-led verbalization and structured judgments or ratings**. Furthermore, **F7** **specific elicitation methods are predominantly employed with certain AI systems (autonomous and driver-assist systems, and predictive and explainable AI systems), possibly influenced by community-specific practices**. In particular, mental models of autonomous and driver-assist systems are predominantly elicited with surveys resembling knowledge tests, emphasizing correctness over unconstrained inferences. Mental models of predictive and explainable AI systems are predominantly elicited with prediction or decision tasks, typically asking participants to predict future systems' states or outcomes. These findings were derived from the extraction and analysis of papers' elicitation methods, grouped by the overarching typology of elicitation described in Section 2.3. Importantly, most studies combine several approaches, typically two (35%) or three (23%), and only eleven studies (13%) relied on a single elicitation approach according to our coding scheme. Figure 7 summarizes the prevalence of elicitation techniques employed, grouped by categories.

Researcher-led verbalizations (66%) and structured judgments and ratings (65%) dominate the corpus. Within researcher-led verbalizations, open-ended questions embedded in surveys (39%) and interviews (34%) are the most common. Within structured judgments and ratings, the most common elicitation techniques include agreement or self-reported comprehension scales (38%), prediction or decision tasks (23%), and vignette or scenario-based questionnaires (22%). Among these elicitation methods, prediction/decision tasks and vignette/scenario-based questionnaires echo the inferential feature of mental models because they ask participants to predict future system states from unseen inputs or scenarios.

Participant-led verbalizations appear less frequently and mainly include retrospective commentary (25%) and think-aloud protocol (17%), aimed to probe reasoning retrospectively (e.g., from a video recording) or as it unfolds. Finally, artifact-based elicitations remain rare, with free drawing (11%) and card-based techniques (5%) being the most prevalent techniques. Artifact-based techniques such as drawing seem particularly employed to externalize broader relationships between AI systems, participants, and the environment [83].

The relational analysis reveals that studies on autonomous or driving-assisting vehicles significantly rely on surveys resembling knowledge or comprehension tests ($p_v < 0.01$), an approach that emphasizes mental models' correctness over their structure. This approach risks framing mental models as the recall of disconnected factual knowledge, overlooking their causal and inferential nature. This relationship also suggests that specific research communities develop their own repertoires of elicitation methods and instruments, which are sometimes reused across studies. For example, the survey developed by Beggiato et al. [12] was reused and adapted in three subsequent studies on autonomous and driving-assistance systems [16, 55, 114]. Conversely, studies on predictive and explainable AI systems for decision-making significantly rely on prediction or decision tasks ($p_v < 0.01$), i.e., asking participants to anticipate systems' future states. This methodological alignment is unsurprising, as predictive systems naturally lend themselves to prediction-based tasks, offering a quantifiable measure of mental model accuracy. While the overall association between elicitation methods and epistemic assumptions (functional, structural, or analogical/anthropomorphic) was significant ($p_v < 0.05$), no specific pairwise relationship reached significance after correction. This suggests that while researchers' methodological choices may not be unrelated to their epistemic assumptions, no consistent pairing emerges across the literature.

Finally, a large majority (88%) of studies compared mental models elicited against an assumed ground truth, across experimental

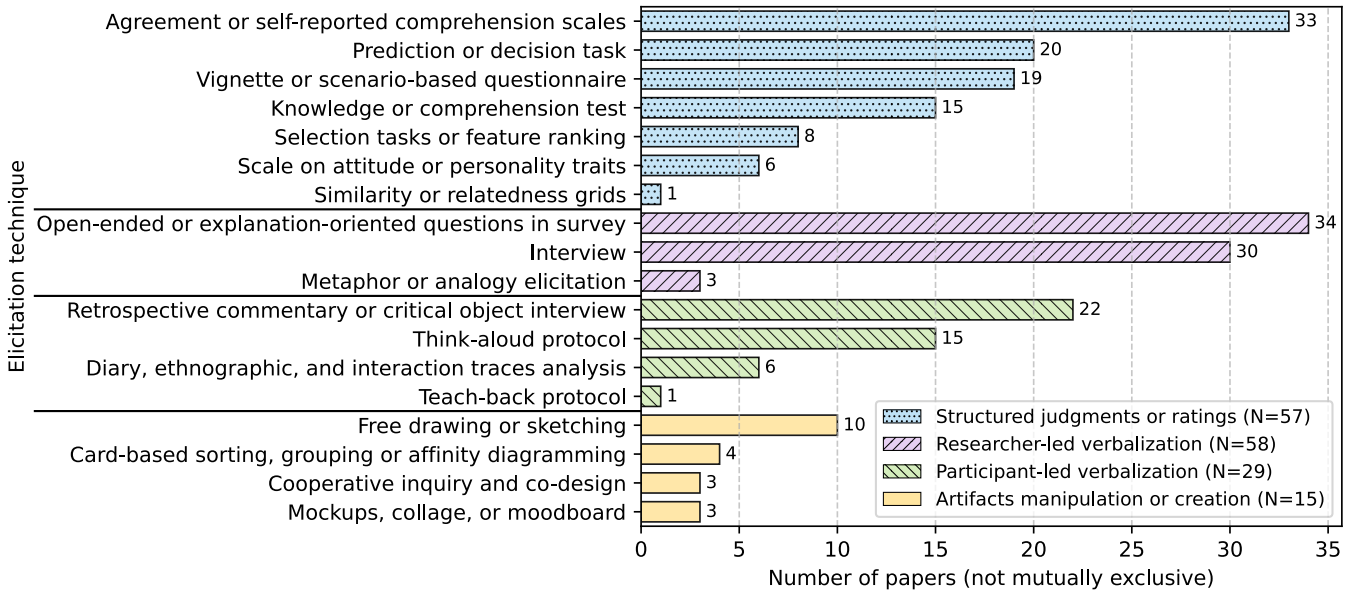


Figure 7: Elicitation methods used across the corpus. Number of papers (not mutually exclusive) employing each elicitation technique, grouped by overarching method category from the typology described in Section 2.3.

conditions, or participant groups. This confirms that most studies elicit mental models for evaluative purposes rather than purely descriptive or exploratory inquiries. About 39% of the studies explicitly examined changes in mental models over time, typically through repeated elicitation before and after exposure to an AI system. Such an experimental design emphasizes the dynamic nature of mental models.

Overall, these findings align with earlier observations that mental model elicitation in HCI is diverse and often triangulated [54]. However, findings reveal that elicitation methods are unevenly used across types of AI systems. Autonomous and driver-assisting vehicles predominantly rely on knowledge and comprehension tests, probably reflecting community-specific practices. Predictive and explainable AI systems for decision-making predominantly naturally invite prediction and decision tasks, i.e., participants predicting unseen or future systems' outputs. These results underscore the opportunity for methodological innovation—approaching the elicitation of mental models in ways that go beyond what community-specific traditions or system affordances naturally invite.

4.4.4 Author-reported limitations. **F8 Author-reported limitations largely reflect common limitations in HCI and provide little insight into the specific challenges of eliciting mental models of AI systems.** We collected limitations reported by the authors in order to understand their perspectives on the limits of mental model elicitation. We extracted and analyzed the limitations explicitly acknowledged in each paper and in relation to the method. Our inductive analysis led to 13 types of limitations and to 4 overarching types of limitations, whose prevalences are depicted in Figure 8.

Several limitations were often reported together in a paper, and are primarily concerned with the methods and protocol (77%), participants (76%), and the system or setup (58%). The most prevalent limitation is about the **limited elicitation scope** (58%), as many studies report capturing only partial aspects of participants' mental models. This challenge is inherent to elicitation, which is largely interpretative and often assumes a certain type of mental model to be elicited. The second most prevalent limitation concerned **limited sampling and representativeness** (50%). While unspecific to mental model research, this problem may be amplified as prior exposure to AI systems varies widely across populations and cultures, affecting mental model construction 2.1. A small but distinctive category of limitations relates to **temporal and external validity**. Laypeople's familiarity with AI systems has changed dramatically within just a few years, given the rapid diffusion of consumer-facing AI services (e.g., ChatGPT in 2022). Consequently, findings from studies eliciting people's mental models of such systems [1] may be short-lived. Together, limitations of sampling, representativeness, and external validity underscore the importance of capturing participants' sociodemographic, prior knowledge, and habitual use of AI systems, given that people's mental models are shaped by background knowledge and evolve with experience. Such limitations also highlight an opportunity to explore longitudinal protocols, which remain underrepresented in our corpus (3%) [11, 12, 21].

Finally, the relational analysis (see Section 3.4) shows that author-reported limitations are not related to any other variables presented earlier—such as the type of AI system studied, the epistemic assumptions about mental models, or the elicitation methods employed. Rather than reflecting concerns specific to mental models or AI systems, these limitations appear to mirror broader issues in HCI or to be idiosyncratic to particular studies. As such, they offer limited

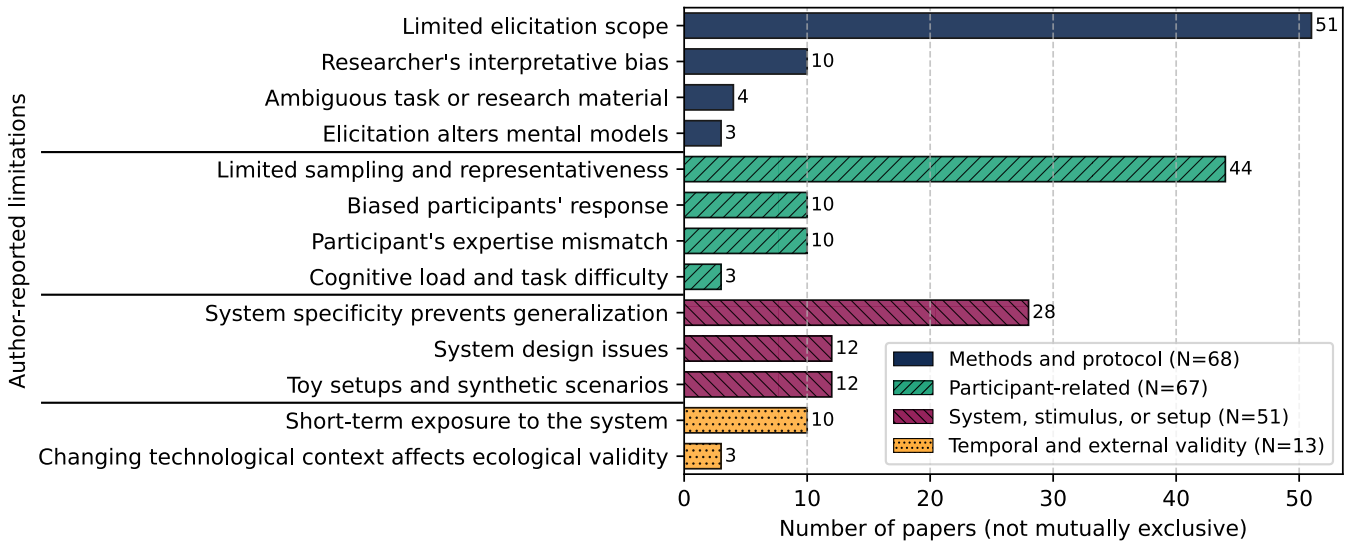


Figure 8: Author-reported limitations related to mental model elicitation. Number of papers (not mutually exclusive) reporting each type of limitation, grouped by broader thematic category.

insights into the specific challenges of eliciting mental models of AI systems.

4.5 Methodological quality, transparency, and participant sampling

This subsection briefly reports on indicators of methodological quality and transparency collected. While not central to our research questions, it remains a common practice in systematic literature reviews. Regarding **participants and recruitment**, over one-third of studies rely on student samples (35%), recruited by convenience (25%) or through internal university channels (36%). This skews participant samples toward younger, more educated individuals from developed countries. These sampling biases are common in psychology and HCI research, and future work would benefit from engaging with more representative samples if deemed relevant for their research questions. Indicators for **transparency and reproducibility** reveal that a minority of studies report[18] being ethically approved (39%), employing reliability psychometrics (24%), or sharing supplementary material (34%). None reported preregistration, even though its benefits for planning, transparency, and reflexivity extend beyond hypothesis-driven experiments. Given that research on mental models of AI systems has expanded since 2019, these findings underscore the need to strengthen recruitment, transparency, and reproducibility practices.

5 Discussion

Section 5.1 synthesizes and discusses the implications of our findings. Then, Section 5.2 provides practical guidelines to support more deliberate and reflective methods for eliciting mental models of AI systems. Finally, Section 5.3 discusses the limitations of our work.

[18]We only checked whether ethics approval, supplementary material, and preregistration were reported in the paper, not whether they were implemented.

5.1 Synthesis and implications

This subsection synthesizes and discusses our main findings in light of our research questions and related work.

*Elicitation practices and embedded epistemic assumptions (**RQ1**).* We found that elicitation methods are typically triangulated, most often drawing from researcher-led verbalizations (e.g., open-ended questions in surveys, interviews) and structured judgments or ratings **F6**. This pattern is consistent with prior reviews of mental model elicitation in HCI [54], and we did not identify elicitation techniques prevalent to human–AI interaction contexts that would not also appear in broader HCI contexts. Furthermore, our results also suggest that mental model elicitation serves different purposes **F3**: (1) system evaluation, typically targeting a specific system or prototype to which participants are exposed during the study, an approach most common in predictive and explainable AI for decision-making papers; (2) understanding shared perceptions of AI in everyday contexts, mainly applied to consumer-oriented AI and voice assistants. The former usage blurs the line between mental models and folk theories, underscoring the need for researchers to clarify which construct best suits their intended purpose. Following our framing that elicitation methods presuppose the type of mental model they make visible (e.g., their epistemic assumption), we observed that the classic functional–structural distinction prevalent in the HCI literature (see Section 2.2) must be extended to include analogical and anthropomorphic lenses **F4**. It is unsurprising that AI systems are often approached through such lenses: robots, agents, and avatars are often designed to be social and invite anthropomorphism. What was previously undocumented, however, is the explicit association between anthropomorphism and the mental model concept. This raises a theoretical question: Is the mental model construct, as defined in cognitive psychology, compatible with analogical and anthropomorphic approaches? We argue that it is, considering that cognitive psychology has examined analogical

reasoning as a mechanism for mental model formation—for example, Gentner and Gentner [42] showed how people reason about electricity through analogical *structure mapping* from unrelated phenomena such as flowing water or teeming crowds, and this mapping enables new causal relationships to be established. In this view, analogies are cognitive mechanisms supporting inference and prediction. Anthropomorphism can similarly be understood as a special case of analogical transfer, where human-human interaction schemata are mapped onto novel human-AI interactions. For instance, perceiving a large language model as sycophantic—overly agreeable or eager to please—may help users reason about how it would function within agentic or automated pipelines, and what kinds of breakdowns might arise. We foresee research opportunities in moving beyond static anthropomorphic traits to examine how anthropomorphism and analogy shape reasoning about AI systems' functions and structures.

*Methodological choices: pragmatic rather than theoretical commitments (**RQ2**).* Consistent with prior research [54], we found that mental model definitions, when provided, often borrow domain-specific influences [100, 101] but rarely capture all distinctive features of mental models as framed in cognitive psychology **F1**. More importantly, definitions appear decoupled from elicitation choices and from the type of AI system studied **F2**. Instead, both elicitation methods and their underlying assumptions relate to the type of AI system investigated **F7** **F5**, particularly in studies of autonomous and driver-assist vehicles, AI-embedded robots, and predictive and explainable AI systems. These patterns suggest that elicitation choices are shaped more by system characteristics and community-specific practices than theoretical commitments. We postulate that papers on autonomous vehicles and driver-assist systems borrow methods from related studies within their immediate research community [12, 16, 55, 114]. A network analysis of papers' citations could further confirm this hypothesis. Predictive and explainable AI systems mainly perform classification, recognition, or detection. Such tasks readily afford measuring accuracy in prediction tasks, e.g., asking participants to guess the system's outcome on a certain input, which could explain the predominant adoption of such elicitation within these papers. Pragmatic over theoretical commitments is not inherently problematic, especially when guided by research comparability, i.e., reusing methods in follow-up papers. However, some methods risk reducing mental models to factual recall, overlooking their inferential and causal nature, and thereby weakening its theoretical strength.

*Are AI properties complicating mental model elicitation? (**RQ3**).* Lastly, we found that author-reported limitations about mental model elicitation largely mirror those found in HCI **F8**. They offer little insight into the specific challenges of eliciting mental models of AI systems. Only a few, such as concerns about temporal and external validity, relate to the rapid mainstreaming of consumer-facing AI [1, 145, 147], which may reshape people's schemata of AI and influence how they reason about AI systems in future studies. This calls for longitudinal research that captures how people's reasoning about AI co-evolves with technological changes. Therefore, our third research question remains unanswered, but could be examined through complementary approaches. For instance, interviewing researchers from our corpus could help better understand methodological challenges specific to the mental model elicitation of AI systems.

5.2 Guidelines for HCI research

According to Hu and Twidale [54], the widening gap between empirical and theoretical work on mental models calls for a cohesive theoretical framework to reconcile divergent perspectives in HCI. However, extensive discussions and refinements on mental model thoery already took place in the 1990s [8, 96, 112, 116, 118] but remain largely ignored in the empirical studies from our corpus. Reinstating mental models as a bridging concept [30] between theory and practice through further theoretical integration therefore risks repeating this limited uptake. We argue that a more effective lever lies in supporting reflective methodological practices that make explicit the theoretical assumptions embedded in how mental models are elicited. Accordingly, we propose nine guidelines that reflect our interpretative stance on the methods reviewed in our corpus, and aim to promote stronger methodological practices on mental models of AI systems.

G1 Report a definition of mental model grounded in its core theoretical attributes. Mental model definitions should reflect the core theoretical attributes surfaced in our corpus: mental models are (1) internal representations of (2) external systems or phenomena; they encode (3) structure and causal relations; (4) enable reasoning and inference; (5) are dynamic, constructed from experience and context; are (6) condensed and often incomplete. We do not prescribe a specific definition but draw the reader's attention to the concise and well-motivated one proposed by Rook [119], which captures these key attributes.

G2 Specify the scope of the mental model. Clarify what participants are expected to form a mental model *about*. In HCI, the scope can range from a specific system (e.g., Alexa Echo), to a class of systems (e.g., voice assistants), or an imagined one. Making this scope explicit helps define the study's focus and the generalizability of its insights.

G3 Specify the type of mental model elicited (epistemic assumption). Elicitation methods presuppose the kind of reasoning they make visible—whether functional (how the system behaves), structural (how it works internally), or analogical and anthropomorphic (how it relates to other entities or human beings). Making these underlying epistemic assumptions explicit can enhance the interpretability and comparability of research.

G4 Assess participants' socio-demographics, prior knowledge, and experience related to the system. Mental models are shaped by schemata, i.e., prototypical knowledge formed by prior experience. Therefore, assessing participants' backgrounds—such as familiarity with similar technologies, domain expertise, and habitual use—contextualizes findings and supports longitudinal comparison. This is particularly important in human-centered AI research, given the rapid evolution of AI technologies and users' increasing familiarity with them.

G5 Favour elicitation methods that force participants to reason and infer. A defining property of mental models is their capacity to support reasoning and inference beyond what is known. To elicit mental models—rather than recalled facts or static beliefs—use methods that compel participants to reason about unseen

situations or guess the system's future states or outcomes. Methods can include prediction tasks (when applicable), as well as "what if" questions (e.g., "What would happen if X?", "What if you had done X instead?") in surveys and interviews, which probe causal reasoning beyond typical use [17].

G6 Consider methods that capture participants' uncertainty. Models are condensed representations, notoriously described as bounded and incomplete, especially when representing complex systems [101]. Therefore, expression of doubt, uncertainty, and ambiguity is not noise, but can be seen as a constitutive element of users' mental models. Methods that capture uncertainty—from confidence ratings in surveys to attending to expressions of doubt when coding interviews—can be valuable, yet are rarely emphasized in mental model empirical research.

G7 Assess whether participants experienced self-reflection during elicitation. Elicitation can be self-reflective for participants, i.e., alter the very mental model it seeks to capture. The way this self-reflection unfolds is difficult to grasp empirically. Asking participants whether the process led them to rethink or clarify their understanding can at least help acknowledge its presence and improve methodological transparency.

G8 Justify triangulation of elicitation methods and report on the integration of heterogeneous insights. The benefits of triangulating elicitation methods—using multiple techniques to elicit the same mental model—are often assumed rather than justified. Clarifying why triangulation is necessary and reporting how and whether insights from different methods are integrated can strengthen methodological transparency and support comparability across research.

G9 Use preregistration for conceptual and methodological commitments. Although often associated with hypothesis-driven experiments, preregistration can promote reflexivity and transparency across all kind of research methods. We argue that it remains the best tool to implement these guidelines, for instance, by stating upfront the mental model's scope, epistemic assumptions, and elicitation approach.

5.3 Limitations

Our work presents three main limitations. First, relevant studies may have been wrongly excluded during the identification and screening phases. Several relevant papers [13, 46, 81, 135] were not initially retrieved during the keyword-based search. One reason may be the incomplete set of AI-related search keywords used; however, Blömacher et al. [13] matches our search strategy, suggesting that keyword generality alone does not fully explain omissions at the identification stage. Other papers may have been missed because they do not use our AI-related keywords [81] or variations of the mental model terminology [46, 135]. We acknowledge that using the wildcard expression `mental model*` instead of `mental model` when possible could have captured Vanderlyn et al. [135] and potentially other relevant studies. The screening of titles and abstracts may also have introduced wrongful exclusions. However, sanity checks against a human annotator showed high sensitivity, suggesting that the automated screening retrieved most papers identified as relevant through manual screening. Prior work further indicates that human screeners may be less sensitive than LLMs for title and abstract screening [19, 64], potentially due to fatigue. As a result, it remains uncertain whether manual screening would have yielded better coverage. Citation snowballing might improve the retrieval of relevant papers, particularly as the literature continues to grow. Second, the scope is limited to studies that use the term "mental model" explicitly, although the concept may be operationalized using other formulations (e.g., "user understanding" or "reasoning"). Including broader terminology could have yielded too many irrelevant results in the identification phase and introduced interpretive challenges in determining conceptual alignment. Third, screening and data extraction required the authors' interpretation, introducing the possibility of systematic bias. These three limitations are largely inherent to systematic literature reviews, and we hope that sharing extensive supplementary material can support future meta-analyses on the topic.

6 Conclusion

This paper investigates how mental models are elicited in Human-AI interaction research. The work is guided by the assumption that epistemological questions about how mental models are elicited are as revealing as ontological debates about what they are. Thus, we analyzed the way researchers elicit mental models, the type of mental model their methods presuppose, and how these vary across definitions used and AI system types. We conducted a systematic literature review of 88 empirical studies eliciting humans' mental models of AI systems. Our analysis shows that (1) the goals of mental model elicitation bifurcate between system-specific evaluation and class-level probes of lay theories; (2) epistemic assumptions extend beyond the classic functional–structural lens (how a system behaves versus how it works internally) to include analogical and anthropomorphic framings; and (3) elicitation methods are shaped more by system characteristics and community-specific practices than theoretical commitments, particularly for predictive and explainable AI systems and autonomous or driver-assist vehicles. Building on these findings, we proposed nine practical guidelines for defining and eliciting mental models of AI systems. Together, they can support the human–AI interaction community by helping to **clarify conceptual grounding**, by encouraging definitions that retain mental models' key theoretical attributes; **promote methodological transparency and comparability**, by making epistemic assumptions, elicitation scope, and triangulation strategies explicit; and **foster reflective and context-aware inquiry**, by attending to uncertainty, participants' self-reflection, and the dynamic nature of human understanding of AI systems. We hope these contributions strengthen conceptual coherence and methodological deliberation around mental models, ultimately supporting more cumulative and transparent research on human–AI interaction.

7 Generative AI disclosure

Beyond editing the author's own text, this systematic literature review employed large language models (LLMs) for screening and data extraction. We used locally hosted LLMs when possible and did not process copyrighted studies when using third-party models. Furthermore, we implemented best practices informed by the research literature, including manual checks of the LLM's output.

We believe that LLM use in systematic reviews should be weighed against the cons of human work, where biases to reduce long and repetitive tasks can also impact research quality.

Acknowledgments

This research is funded by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Actions Postdoctoral Fellowship (HAMLU, Grant Agreement No. 101153987). Prof. Ophelia Deroy is funded by the European Union under the European Innovation Council (EMERGE, Grant Agreement No. 101070918). We thank Xinhui Hu and Prof. Michael Twidale for their helpful discussions on mental models at an early stage of this research. Some figure elements are used under a paid license from The Noun Project.

References

[1] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. 2019. More than smart speakers: security and privacy perceptions of smart home personal assistants. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security* (Santa Clara, CA, USA) *(SOUPS'19)*. USENIX Association, USA, 451–466.

[2] Rudaiba Adnin and Maitraye Das. 2024. "I look at it as the king of knowledge": How Blind People Use and Understand Generative AI Tools. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 64, 14 pages. doi:10.1145/3663548.3675631

[3] Luiz Agner, Barbara Necyk, and Adriano Renzi. 2020. Recommendation Systems and Machine Learning: Mapping the User Experience. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*, Aaron Marcus and Elizabeth Rosenzweig (Eds.). Springer International Publishing, Cham, 3–17.

[4] Joseph W. Alba and Lynn Hasher. 1983. Is memory schematic? *Psychological Bulletin* 93, 2 (1983), 203. https://psycnet.apa.org/record/1983-20271-001 Publisher: American Psychological Association.

[5] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 2 (2020), 1–37. doi:10.1145/3366485

[6] Robert W. Andrews, J. Mason Lilly, Divya Srivastava, and Karen M. Feigh. 2023. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science* 24, 2 (2023), 129–175. doi:10.1080/1463922X.2022.2061080

[7] Tabinda Aziz, Yukio Horiguchi, and Tetsuo Sawaragi. 2014. Towards understanding the development of driver's mental model of a Lane Departure Warning system while driving. *Journal of Advanced Mechanical Design, Systems, and Manufacturing* 8, 3 (2014), JAMDSM0032–JAMDSM0032. doi:10.1299/jamdsm.2014jamdsm0032

[8] Lisanne Bainbridge. 1992. Mental models in cognitive skill: the example of industrial process operation. *Models in the Mind* (1992), 119–143. Publisher: Academic Press London.

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. doi:10.1609/hcomp.v7i1.5285

[10] Piraye Bayman and Richard E. Mayer. 1984. Instructional manipulation of users' mental models for electronic calculators. *International Journal of Man-Machine Studies* 20, 2 (1984), 189–199. Publisher: Elsevier.

[11] Matthias Beggiato and Josef F Krems. 2013. The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation research part F: traffic psychology and behaviour* 18 (2013), 47–57. doi:10.1016/j.trf.2012.12.006

[12] Matthias Beggiato, Marta Pereira, Tibor Petzoldt, and Josef Krems. 2015. Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation research part F: traffic psychology and behaviour* 35 (2015), 75–84. doi:10.1016/j.trf.2015.10.005

[13] Katja Blömacher, Gerhard Nöcker, and Markus Huff. 2020. The evolution of mental models in relation to initial information while driving automated. *Transportation research part F: traffic psychology and behaviour* 68 (2020), 198–217.

[14] Michelle Brachman, Siya Kunde, Sarah Miller, Ana Fucs, Samantha Dempsey, Jamie Jabbour, and Werner Geyer. 2025. Building Appropriate Mental Models: What Users Know and Want to Know about an Agentic AI Chatbot. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 247–264. doi:10.1145/3708359.3712071

[15] Michelle Brachman, Qian Pan, Hyo Jin Do, Casey Dugan, Arunima Chaudhary, James M. Johnson, Priyanshu Rai, Tathagata Chakraborti, Thomas Gschwind, Jim A Laredo, Christoph Miksovic, Paolo Scotton, Kartik Talamadupula, and Gegi Thomas. 2023. Follow the Successful Herd: Towards Explanations for Improved Use and Mental Models of Natural Language Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 220–239. doi:10.1145/3581641.3584088

[16] Claudia Buchner, Chantal Himmels, Jan Schmitz, Tanja Stoll, and Martin Baumann. 2024. Exploring Urban Challenges: Understanding Advanced Driver Assistance Systems in Different Situational Contexts. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Stanford, CA, USA) *(AutomotiveUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3640792.3675709

[17] Ruth MJ Byrne. 2002. Mental models and counterfactual thoughts about what might have been. *Trends in cognitive sciences* 6, 10 (2002), 426–431. https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(02)01974-5?large_figure=true Publisher: Elsevier.

[18] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. doi:10.1145/3359206

[19] Christian Cao, Jason Sang, Rohit Arora, Robbie Kloosterman, Matt Cecere, Jaswanth Gorla, Richard Saleh, David Chen, Ian Drennan, Bijan Teja, Michael Fehlings, Paul Ronksley, Alexander A Leung, Dany E Weisz, Harriet Ware, Mairead Whelan, David B Emerson, Rahul Arora, and Niklas Bobrovitz. 2024. Prompting is all you need: LLMs for systematic review screening. *medRxiv* (2024). doi:10.1101/2024.06.01.24308323

[20] Kathleen Carley and Michael Palmquist. 1992. Extracting, representing, and analyzing mental models. *Social forces* 70, 3 (1992), 601–636. https://academic.oup.com/sf/article-abstract/70/3/601/2232444 Publisher: The University of North Carolina Press.

[21] Cher Carney, John G Gaspar, and William J Horrey. 2022. Longer-term exposure vs training: Their effect on drivers' mental models of ADAS technology. *Transportation research part F: traffic psychology and behaviour* 91 (2022), 329–345.

[22] John M. Carroll and Judith Reitman Olson. 1988. Chapter 2 - Mental Models in Human-Computer Interaction. In *Handbook of Human-Computer Interaction*, MARTIN HELANDER (Ed.). North-Holland, Amsterdam, 45–65. doi:10.1016/B978-0-444-70536-5.50007-5

[23] Weihao Chen, Yuanchun Shi, Yukun Wang, Weinan Shi, Meizhu Chen, Cheng Gao, Yu Mei, Yeshuang Zhu, Jinchao Zhang, and Chun Yu. 2025. Investigating Context-Aware Collaborative Text Entry on Smartphones using Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1084, 20 pages. doi:10.1145/3706598.3713944

[24] Myra Cheng, Angela Y. Lee, Kristina M. Rapuano, Kate Niederhoffer, Alex Liebscher, and Jeffrey Hancock. 2025. From tools to thieves: Measuring and understanding public perceptions of AI through crowdsourced metaphors. (2025). doi:10.48550/arxiv.2501.18045

[25] Yi-Shyuan Chiang, Ruei-Che Chang, Yi-Lin Chuang, Shih-Ya Chou, Hao-Ping Lee, I-Ju Lin, Jian-Hua Jiang Chen, and Yung-Ju Chang. 2020. Exploring the Design Space of User-System Communication for Smart-home Routine Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376501

[26] Jiyong Cho and Jeongyun Heo. 2020. The More You Know, The More You Can Trust: Drivers' Understanding of the Advanced Driver Assistance System. In *HCI in Mobility, Transport, and Automotive Systems. Automated Driving and In-Vehicle Experience Design*, Heidi Krömker (Ed.). Springer International Publishing, Cham, 230–248.

[27] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. doi:10.1145/3397481.3450644

[28] Meghan Clark, Mark W. Newman, and Prabal Dutta. 2017. Devices and Data and Agents, Oh My: How Smart Home Abstractions Prime End-User Mental Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 44 (Sept. 2017), 26 pages. doi:10.1145/3132031

[29] Nancy J. Cooke and Anna L. Rowe. 1994. Evaluating Mental Model Elicitation Methods. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 38, 4 (Oct. 1994), 261–265. doi:10.1177/154193129403800416

[30] Peter Dalsgaard and Christian Dindler. 2014. Between theory and practice: bridging concepts in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*.

Association for Computing Machinery, New York, NY, USA, 1635–1644. doi:10.1145/2556288.2557342

[31] Giuseppe Desolda, Andrea Esposito, Florian Müller, and Sebastian Feger. 2023. Digital Modeling for Everyone: Exploring How Novices Approach Voice-Based 3D Modeling. In *Human-Computer Interaction – INTERACT 2023*, José Abdelnour Nocera, Marta Kristín Lárusdóttir, Helen Petrie, Antonio Piccinno, and Marco Winckler (Eds.). Springer Nature Switzerland, Cham, 133–155.

[32] Hyo Jin Do, Michelle Brachman, Casey Dugan, Qian Pan, Priyanshu Rai, James M. Johnson, and Roshni Thawani. 2024. Evaluating What Others Say: The Effect of Accuracy Assessment in Shaping Mental Models of AI Systems. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 373 (Nov. 2024), 26 pages. doi:10.1145/3686912

[33] K. Anders Ericsson and Herbert A. Simon. 1998. How to Study Thinking in Everyday Life: Contrasting Think-Aloud Protocols With Descriptions and Explanations of Thinking. *Mind, Culture, and Activity* 5, 3 (July 1998), 178–186. doi:10.1207/s15327884mca0503_3

[34] Margarita Esau, Veronika Krauß, Dennis Lawo, and Gunnar Stevens. 2022. Losing Its Touch: Understanding User Perception of Multimodal Interaction and Smart Assistance. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1288–1299. doi:10.1145/3532106.3533455

[35] Lotte Ewals, Lynn J J Heesterbeek, Bin Yu, Kasper van der Wulp, Dimitrios Mavroeidis, Mathias Funk, Chris Snijders, Igor Jacobs, Joost Nederend, and Jon R. Pluyter. 2024. The Impact of Expectation Management and Model Transparency on Radiologists' Trust and Utilization of AI Recommendations for Lung Nodule Assessment on Computed Tomography: Simulated Use Study. *JMIR AI* 3 (13 Mar 2024), e52211. doi:10.2196/52211

[36] Cindy Farris. 2015. The teach back method. *Home healthcare now* 33, 6 (2015), 344–345. https://journals.lww.com/homehealthcarenurseonline/fulltext/2015/06000/the_teach_back_method.13.aspx Publisher: LWW.

[37] Sophie Feinauer, Irene Groh, and Tibor Petzoldt. 2025. The Impact of a Priori Information on Drivers' Mental Models, Attitudes, and Behavior in Interaction with Partial and Conditional Driving Automation. *International Journal of Human–Computer Interaction* 41, 7 (2025), 3828–3840. doi:10.1080/10447318.2024.2341190

[38] KJ Feng, David W McDonald, and Amy X Zhang. 2025. Levels of Autonomy for AI Agents. *arXiv preprint arXiv:2506.12469* (2025). doi:10.48550/arXiv.2506.12469

[39] Center for Security, Emerging Technology, and Our World in Data. 2025. Annual Scholarly Publications on Artificial Intelligence. Our World in Data (processed data). https://archive.ourworldindata.org/20250909-093708/grapher/annual-scholarly-publications-on-artificial-intelligence.html?tab=line Data from CSET and processed by Our World in Data..

[40] Francisco Javier Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models. In *Proceedings of the 11th International Conference on Natural Language Generation*, Emiel Krahmer, Albert Gatt, and Martijn Goudbeek (Eds.). Association for Computational Linguistics, Tilburg University, The Netherlands, 99–108. doi:10.18653/v1/W18-6511

[41] Pedro García García, Enrico Costanza, Jhim Verame, Diana Nowacka, and Sarvapali D. Ramchurn. 2021. Seeing (Movement) is Believing: The Effect of Motion on Perception of Automatic Systems Performance. *Human–Computer Interaction* 36, 1 (2021), 1–51. doi:10.1080/07370024.2018.1453815

[42] Dedre Gentner and Donald R. Gentner. 1983. Flowing waters or teeming crowds: Mental models of electricity. *Mental models* 99 (1983), 129.

[43] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376316

[44] Kimberly Glasgow, Jonathon Kopecky, John Gersh, and Adam C. G. Crego. 2022. Won't you see my neighbor? User predictions, mental models, and similarity-based explanations of AI classifiers. abs/2201.13301 (2022). doi:10.54941/ahfe1001440

[45] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of management annals* 14, 2 (2020), 627–660.

[46] G Mark Grimes, Ryan M Schuetzler, and Justin Scott Giboney. 2021. Mental models and expectation violations in conversational AI interactions. *Decision Support Systems* 144 (2021), 113515.

[47] Nader Hanna and Deborah Richards. 2016. Do birds of a feather work better together? The impact of a match in personality between humans and virtual agents on a shared mental model during collaboration. *Int. J. Comput. Intell. Stud.* 5, 2 (Jan. 2016), 162–179. doi:10.1504/IJCISTUDIES.2016.077134

[48] Nader Hanna and Deborah Richards. 2018. The impact of multimodal communication on a shared mental model, trust, and commitment in human–intelligent virtual agent teams. *Multimodal Technologies and Interaction* 2, 3 (2018), 48. doi:10.3390/mti2030048

[49] Wu He, Sanda Erdelez, and Feng-Kwei Wang. 2010. Examining a case-based reasoning retrieval system using mental models as a framework. *International Journal of Learning Technology* 5, 1 (2010), 63–79. doi:10.1504/IJLT.2010.031616

[50] Wu He, Sanda Erdelez, Feng-Kwei Wang, and Chi-Ren Shyu. 2008. The effects of conceptual description and search practice on users' mental models and information seeking in a case-based reasoning retrieval system. *Inf. Process. Manage.* 44, 1 (Jan. 2008), 294–309. doi:10.1016/j.ipm.2007.03.008

[51] Ziyao He, Yunpeng Song, Shurui Zhou, and Zhongmin Cai. 2023. Interaction of Thoughts: Towards Mediating Task Assignment in Human-AI Cooperation with a Capability-Aware Shared Mental Model. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 353, 18 pages. doi:10.1145/3544548.3580983

[52] Alexander Heimerl, Katharina Weitz, Tobias Baur, and Elisabeth André. 2022. Unraveling ML Models of Emotion With NOVA: Multi-Level Explainable AI for Non-Experts. *IEEE Transactions on Affective Computing* 13, 3 (July 2022), 1155–1167. doi:10.1109/TAFFC.2020.3043603

[53] Aike C. Horstmann, Clara Strathmann, Lea Lambrich, and Nicole C. Krämer. 2023. Alexa, What's Inside of You: A Qualitative Study to Explore Users' Mental Models of Intelligent Voice Assistants. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (Würzburg, Germany) *(IVA '23)*. Association for Computing Machinery, New York, NY, USA, Article 5, 10 pages. doi:10.1145/3570945.3607335

[54] Xinhui Hu and Michael Twidale. 2023. A Scoping Review of Mental Model Research in HCI from 2010 to 2021. In *HCI International 2023 – Late Breaking Papers*, Masaaki Kurosu, Ayako Hashizume, Aaron Marcus, Elizabeth Rosenzweig, Marcelo M. Soares, Don Harris, Wen-Chin Li, Dylan D. Schmorrow, Cali M. Fidopiastis, and Pei-Luen Patrick Rau (Eds.). Springer Nature Switzerland, Cham, 101–125.

[55] Chunxi Huang, Dengbo He, Xiao Wen, and Song Yan. 2023. Beyond adaptive cruise control and lane centering control: drivers' mental model of and trust in emerging ADAS technologies. *Frontiers in Psychology* Volume 14 - 2023 (2023). doi:10.3389/fpsyg.2023.1236062

[56] Apoorva Hungund, Ganesh Pai, and Anuj K. Pradhan. 2024. Does training improve users' mental models about adaptive cruise control? *Traffic Safety Research* 6 (Jan. 2024), e000041. doi:10.55329/aqze5695

[57] Aleksi Huotala, Miikka Kuutila, Paul Ralph, and Mika Mäntylä. 2024. The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*. ACM, Salerno Italy, 262–271. doi:10.1145/3661167.3661172

[58] Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press.

[59] Philip N. Johnson-Laird. 2004. The history of mental models. In *Psychology of reasoning*. Psychology Press, 189–222. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203506936-10/history-mental-models-johnson-laird

[60] Philip N. Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences* 107, 43 (Oct. 2010), 18243–18250. doi:10.1073/pnas.1012933107

[61] Philip N Johnson-Laird and Fabien Savary. 1999. Illusory inferences: A novel class of erroneous deductions. *Cognition* 71, 3 (1999), 191–229.

[62] P. N. Johnson-Laird, Paolo Legrenzi, and Maria Sonino Legrenzi. 1972. Reasoning and a sense of reality. *British Journal of Psychology* 63, 3 (Aug. 1972), 395–400. doi:10.1111/j.2044-8295.1972.tb01287.x

[63] Natalie A. Jones, Helen Ross, Timothy Lynam, Pascal Perez, and Anne Leitch. 2011. Mental Models: An Interdisciplinary Synthesis of Theory and Methods. *Ecology and Society* 16, 1 (2011). doi:10.5751/ES-03802-160146

[64] Lucas Joos, Daniel A. Keim, and Maximilian T. Fischer. 2025. Cutting Through the Clutter: The Potential of LLMs for Efficient Filtration in Systematic Literature Reviews. *EuroVis Workshop on Visual Analytics (EuroVA) 2025* (2025). doi:10.2312/EUROVA.20251105

[65] Riitta Jääskeläinen. 2010. Think-aloud protocol. In *Handbook of Translation Studies*, Yves Gambier and Luc Van Doorslaer (Eds.). Vol. 1. John Benjamins Publishing Company, Amsterdam, 371–373. doi:10.1075/hts.1.thi1 Series Title: Handbook of Translation Studies.

[66] Daniel Kats and Mahmood Sharif. 2022. "I Have No Idea What a Social Bot Is": On Users' Perceptions of Social Bots and Ability to Detect Them. In *Proceedings of the 10th International Conference on Human-Agent Interaction* (Christchurch, New Zealand) *(HAI '22)*. Association for Computing Machinery, New York, NY, USA, 32–40. doi:10.1145/3527188.3561928

[67] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1723–1734. doi:10.1145/3593013.3594111

[68] Willett Kempton. 1986. Two Theories of Home Heat Control*. *Cognitive Science* 10, 1 (Jan. 1986), 75–90. doi:10.1207/s15516709cog1001_3

[69] Qusai Khraisha, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. 2024. Can large language models replace humans in systematic reviews? Evaluating GPT's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods* 15, 4 (July 2024), 616–626. doi:10.1002/jrsm.1715

[70] Mohammad Kianpisheh, Alex Mariakakis, and Khai N. Truong. 2024. exHAR: An Interface for Helping Non-Experts Develop and Debug Knowledge-based Human Activity Recognition Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 11 (March 2024), 30 pages. doi:10.1145/3643500

[71] David Kieras and Peter G. Polson. 1985. An approach to the formal analysis of user complexity. *International journal of man-machine studies* 22, 4 (1985), 365–394.

[72] David E. Kieras and Susan Bovair. 1984. The role of a mental model in learning to operate a device. *Cognitive Science* 8, 3 (1984), 255–273. doi:10.1016/S0364-0213(84)80003-8

[73] Bart P Knijnenburg and Martijn C Willemsen. 2016. Inferring capabilities of intelligent agents from their external traits. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 4 (2016), 1–25. doi:10.1145/2963106

[74] Martin Krampell, Ignacio Solís-Marcos, and Magnus Hjälmdahl. 2020. Driving automation state-of-mind: Using training to instigate rapid mental model development. *Applied ergonomics* 83 (2020), 102986. doi:10.1016/j.apergo.2019.102986

[75] Benjamin Krarup, Felix Lindner, Senka Krivic, and Derek Long. 2022. Understanding a Robot's Guiding Ethical Principles via Automatically Generated Explanations. *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)* (Aug. 2022), 627–632. doi:10.1109/CASE49997.2022.9926613

[76] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/2207676.2207678 event-place: Austin, Texas, USA.

[77] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. doi:10.1109/VLHCC.2013.6645235

[78] Johannes Kunkel, Thao Ngo, Jürgen Ziegler, and Nicole Krämer. 2021. Identifying Group-Specific Mental Models of Recommender Systems: A Novel Quantitative Approach. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer International Publishing, Cham, 383–404.

[79] Janice Langan-Fox, Sharon Code, and Kim Langfield-Smith. 2000. Team Mental Models: Techniques, Methods, and Analytic Approaches. *Human Factors* 42, 2 (2000), 242–271. doi:10.1518/001872000779656534 PMID: 11022883.

[80] Janice Langan-Fox, Anthony Wirth, Sharon Code, Kim Langfield-Smith, and Andrew Wirth. 2001. Analyzing shared and team mental models. *International Journal of Industrial Ergonomics* 28, 2 (2001), 99–112. doi:10.1016/S0169-8141(01)00016-6

[81] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (Nov. 2018), 31 pages. doi:10.1145/3274371

[82] Sunok Lee, Minji Cho, and Sangsu Lee. 2020. What If Conversational Agents Became Invisible? Comparing Users' Mental Models According to Physical Entity of AI Speaker. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 88 (Sept. 2020), 24 pages. doi:10.1145/3411840

[83] Sunok Lee and Sangsu Lee. 2022. "Hey Alexa, Where Are You?" A Drawing Study to Understand Users' Mental Models of the Environments Surrounding Conversational Agents. In *[ ] With Design: Reinventing Design Modes*, Gerhard Bruyns and Huaxin Wei (Eds.). Springer Nature Singapore, Singapore, 2001–2016.

[84] Liping Li, Hsinwen Chang, Weihan Sun, Jin Guo, and Jianchao Gao. 2020. How Drivers Categorize ADAS Functions. In *Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments*, Pei-Luen Patrick Rau (Ed.). Springer International Publishing, Cham, 606–615. doi:10.1007/978-3-030-49788-0_46

[85] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2119–2128. doi:10.1145/1518701.1519023

[86] Shuai Ma, Mingfei Sun, and Xiaojuan Ma. 2023. Modeling Adaptive Expression of Robot Learning Engagement and Exploring Its Effects on Human Teachers. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 70 (Sept. 2023), 48 pages. doi:10.1145/3571813

[87] Edith Maier and Ulrich Reimer. 2023. Analysing the Use of Voice Assistants in Domestic Settings Through the Lens of Activity Theory. In *Human-Computer Interaction*, Masaaki Kurosu and Ayako Hashizume (Eds.). Springer Nature Switzerland, Cham, 366–379.

[88] Kannan Mani and Philip N. Johnson-Laird. 1982. The mental representation of spatial descriptions. *Memory & Cognition* 10, 2 (1982), 181–187. https://link.springer.com/article/10.3758/BF03209220 Publisher: Springer.

[89] Kenneth I. Manktelow. 1981. Recent developments in research on Wason's selection task. *Current Psychological Reviews* 1, 3 (1981), 257–268. Publisher: Springer.

[90] Marijn Martens, Ralf De Wolf, Bettina Berendt, and Lieven De Marez. 2023. Decoding Algorithms: Exploring End-Users' Mental Models of the Inner Workings of Algorithmic News Recommenders. *Digital Journalism* 11, 1 (2023), 203–225. doi:10.1080/21670811.2022.2129402

[91] Aqueasha Martin-Hammond, Sravani Vemireddy, and Kartik Rao. 2019. Exploring Older Adults' Beliefs About the Use of Intelligent Assistants for Consumer Health Information Management: A Participatory Design Study. *JMIR Aging* 2, 2 (11 Dec 2019), e15381. doi:10.2196/15381

[92] Gerald Matthews, Jinchao Lin, April Rose Panganiban, and Michael D. Long. 2020. Individual Differences in Trust in Autonomous Robots: Implications for Transparency. *IEEE Transactions on Human-Machine Systems* 50, 3 (June 2020), 234–244. doi:10.1109/THMS.2019.2947592

[93] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. 2022. GANterfactual—Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning. *Frontiers in Artificial Intelligence* Volume 5 - 2022 (2022), 825565. doi:10.3389/frai.2022.825565

[94] Christian Meske and Enrico Bunde. 2022. Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. *Information Systems Frontiers* 25, 2 (March 2022), 743–773. doi:10.1007/s10796-021-10234-5

[95] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric Ragan. 2021. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. *Proceedings of the International AAAI Conference on Web and Social Media* 15, 1 (May 2021), 421–431. doi:10.1609/icwsm.v15i1.18072

[96] Neville Moray. 1990. A lattice theory approach to the structure of mental models. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 327, 1241 (April 1990), 577–583. doi:10.1098/rstb.1990.0100

[97] Michael A. Nees and Claire Liu. 2022. Mental models of driver monitoring systems: Perceptions of monitoring capabilities in an online U.S-based sample. *Transportation Research Part F: Traffic Psychology and Behaviour* 91 (2022), 484–498. doi:10.1016/j.trf.2022.10.012

[98] Michael A. Nees, Nithya Sharma, and Karli Herwig. 2020. Some Characteristics of Mental Models of Advanced Driver Assistance Systems: A Semi-structured Interviews Approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64, 1 (2020), 1313–1317. doi:10.1177/1071181320641314

[99] Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. 2020. Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 183–191. doi:10.1145/3340631.3394841

[100] D.A. Norman. 1988. *The Psychology of Everyday Things*. Basic Books.

[101] D. A. Norman. 1987. *Some observations on mental models*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 241–244.

[102] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. doi:10.1145/3397481.3450639

[103] Mahsan Nourani, Chiradeep Roy, Jeremy E. Block, Donald R. Honeycutt, Tahrima Rahman, Eric D. Ragan, and Vibhav Gogate. 2022. On the Importance of User Backgrounds and Impressions: Lessons Learned from Interactive AI Applications. *ACM Trans. Interact. Intell. Syst.* 12, 4, Article 28 (Dec. 2022), 29 pages. doi:10.1145/3531066

[104] Fjollë Novakazi, Julia Orlovska, Lars-Ola Bligård, and Casper Wickman. 2020. Stepping over the threshold linking understanding and usage of Automated Driver Assistance Systems (ADAS). *Transportation Research Interdisciplinary Perspectives* 8 (2020), 100252. doi:10.1016/j.trip.2020.100252

[105] Oecd. 2022. *OECD Framework for the Classification of AI systems*. OECD Digital Economy Papers 323. OECD Publishing. doi:10.1787/cb6d9eca-en

[106] Stine Ordemann, Marita Skjuve, Asbjørn Følstad, and Cato Alexander Bjørkli. 2021. Understanding How Chatbots Work: an Exploratory Study of Mental Models in Customer Service Chatbots. *IADIS International Journal on WWW/Internet* 19, 1 (2021). doi:10.33965/ijwi_202119102

[107] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan—a web and mobile app for systematic reviews. *Systematic reviews* 5, 1 (2016), 210.

[108] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan

Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372 (2021). doi:10.1136/bmj.n71

[109] Dasom Park and Kiechan Namkung. 2021. Exploring users' mental models for anthropomorphized voice assistants through psychological approaches. *Applied Sciences* 11, 23 (2021), 11147. doi:10.3390/app112311147

[110] Devasena Pasupuleti, Sreejith Sasidharan, Gayathri Manikutty, Anand M. Das, Praveen Pankajakshan, and Sidney Strauss. 2023. Co-designing the Embodiment of a Minimalist Social Robot to Encourage Hand Hygiene Practises Among Children in India. *International Journal of Social Robotics* 15, 2 (Feb. 2023), 345–367. doi:10.1007/s12369-023-00969-3

[111] STEPHEN J. PAYNE. 1991. A descriptive study of mental models. *Behaviour & Information Technology* 10, 1 (1991), 3–21. doi:10.1080/01449299108924268

[112] Stephen J. Payne. 2003. CHAPTER 6 - Users' Mental Models: The Very Ideas. In *HCI Models, Theories, and Frameworks*, John M. Carroll (Ed.). Morgan Kaufmann, San Francisco, 135–156. doi:10.1016/B978-155860808-5/50006-X

[113] Leah Perlmutter, Eric Kernfeld, and Maya Cakmak. 2016. Situated Language Understanding with Human-like and Visualization-Based Transparency. (2016). doi:10.15607/RSS.2016.XII.040

[114] Linwei Qiao, Jiaqian Li, and Tingru Zhang. 2024. Impacts of Training Methods and Experience Types on Drivers' Mental Models and Driving Performance. In *HCI in Mobility, Transport, and Automotive Systems: 6th International Conference, MobiTAS 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part I* (Washington DC, USA). Springer-Verlag, Berlin, Heidelberg, 44–56. doi:10.1007/978-3-031-60477-5_4

[115] Vivek Rao, Elisa Kwon, and Kosa Goucher-Lambert. 2023. "Like a Moodboard, But More Interactive": The Role of Expertise in Designers' Mental Models and Speculations on an Intelligent Design Assistant. In *Design Computing and Cognition'22*, John S. Gero (Ed.). Springer International Publishing, Cham, 749–765.

[116] Jens Rasmussen. 1979. *On the structure of knowledge - a morphology of metal models in a man-machine system context.* Number 2192 in Risø-M. Risø National Laboratory.

[117] Kirsten M.A. Revell and Neville A. Stanton. 2012. Models of models: filtering and bias rings in depiction of knowledge structures and their implications for design. *Ergonomics* 55, 9 (Sept. 2012), 1073–1092. doi:10.1080/00140139.2012.692818

[118] Yvonne Rogers, Andrew Rutherford, and Peter Bibby. 1992. Models in the mind: Theory, perspective and application. *(No Title)* (1992). https://cir.nii.ac.jp/crid/1130000795201182464

[119] Laura Rook. 2013. Mental models: A robust definition. *The learning organization* 20, 1 (2013), 38–47. https://www.emerald.com/insight/content/doi/10.1108/09696471311288519/ Publisher: Emerald Group Publishing Limited.

[120] William B. Rouse and Nancy M. Morris. 1986. On looking into the black box: Prospects and limits in the search for mental models. 100, 3 (1986), 349. Publisher: American Psychological Association.

[121] David E. Rumelhart. 1984. Schemata and the cognitive system. (1984). https://psycnet.apa.org/record/2011-28549-005 Publisher: Lawrence Erlbaum Associates Publishers.

[122] Juliano Efson Sales, André Freitas, and Siegfried Handschuh. 2020. A User-centred Analysis of Explanations for a Multi-component Semantic Parser. In *Natural Language Processing and Information Systems*, Elisabeth Métais, Farid Meziane, Helmut Horacek, and Philipp Cimiano (Eds.). Springer International Publishing, Cham, 37–44.

[123] Rohan Sanghera, Arun James Thirunavukarasu, Marc El Khoury, Jessica O'Logbon, Yuqing Chen, Archie Watt, Mustafa Mahmood, Hamid Butt, George Nishimura, and Andrew AS Soltan. 2025. High-performance automated abstract screening with large language model ensembles. *Journal of the American Medical Informatics Association* 32, 5 (2025), 893–904. https://academic.oup.com/jamia/article-abstract/32/5/893/8090299 Publisher: Oxford University Press.

[124] Maarten Schadd, Tjeerd Schoonderwoerd, Karel Van den Bosch, Olaf H. Visker, Tjalling Haije, and Kim H. Veltman. 2022. "I'm Afraid I Can't Do That, Dave"; Getting to Know Your Buddies in a Human–Agent Team. *Systems* 10, 1 (2022), 15. doi:10.3390/systems10010015

[125] Roger C. Schank and Robert P. Abelson. 1975. Scripts, plans, and knowledge. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1* (Tblisi, USSR) *(IJCAI'75)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 151–157.

[126] Ruben Schlagowski, Maurizio Volanti, Katharina Weitz, Silvan Mertes, Johanna Kuch, and Elisabeth André. 2024. The feeling of being classified: raising empathy and awareness for AI bias through perspective-taking in VR. *Frontiers in Virtual Reality* Volume 5 - 2024 (2024), 1340250. doi:10.3389/frvir.2024.1340250

[127] Jan Maarten Schraagen, Pia Elsasser, Hanna Fricke, Marleen Hof, and Fabyen Ragalmuto. 2020. Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64, 1 (2020), 339–343. doi:10.1177/1071181320641077

[128] Claire M. Segijn, Joanna Strycharz, Anna Turner, and Suzanna J. Opree. 2024. Conversation-Related Advertising and Electronic Eavesdropping: Mapping Perceptions of Phones Listening for Advertising in the United States, the Netherlands, and Poland. *Social Media + Society* 10, 4 (Oct. 2024), 20563051241288448. doi:10.1177/20563051241288448 Publisher: SAGE Publications Ltd.

[129] Herbert A Simon et al. 1972. Theories of bounded rationality. *Decision and organization* 1, 1 (1972), 161–176.

[130] Adarsa Sivaprasad, Ehud Reiter, Nava Tintarev, and Nir Oren. 2024. Evaluation of Human-Understandability of Global Model Explanations Using Decision Tree. In *Artificial Intelligence. ECAI 2023 International Workshops*. Springer Nature Switzerland, Cham, 43–65.

[131] Fran Soljacic, Theresa Law, Meia Chita-Tegmark, and Matthias Scheutz. 2024. Robots in healthcare as envisioned by care professionals. *Intelligent Service Robotics* 17, 3 (March 2024), 685–701. doi:10.1007/s11370-024-00523-8

[132] Vaynee Sungeelee, Nathanaël Jarrassé, Téo Sanchez, and Baptiste Caramiaux. 2024. Comparing Teaching Strategies of a Machine Learning-based Prosthetic Arm. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 715–730. doi:10.1145/3640543.3645170

[133] Nathan L. Tenhundfeld, Hannah M. Barr, Emily H. O'Hear, and Kristin Weger. 2022. Is My Siri the Same as Your Siri? An Exploration of Users' Mental Model of Virtual Personal Assistants, Implications for Trust. *IEEE Transactions on Human-Machine Systems* 52, 3 (June 2022), 512–521. doi:10.1109/THMS.2021.3107493

[134] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. 2007. How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 31–40. doi:10.1145/1240624.1240630

[135] Lindsey Morgan Vanderlyn, Dirk Väth, and Thang Vu. 2025. Understanding the Role of Mental Models in User Interaction with an Adaptive Dialog Agent. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 989–1015.

[136] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–39. doi:10.1145/3476068

[137] Jennifer Villareale, Casper Harteveld, and Jichen Zhu. 2022. "I Want To See How Smart This AI Really Is": Player Mental Model Development of an Adversarial AI Player. *Proc. ACM Hum.-Comput. Interact.* 6, CHI PLAY, Article 219 (Oct. 2022), 26 pages. doi:10.1145/3549482

[138] Sarah E. Walsh and Karen M. Feigh. 2023. Mental Models of AI Performance and Bias of Nontechnical Users. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 4116–4121. doi:10.1109/SMC53992.2023.10393994

[139] P. C. Wason. 1968. Reasoning about a Rule. *Quarterly Journal of Experimental Psychology* 20, 3 (Aug. 1968), 273–281. doi:10.1080/14640746808400161

[140] Katharina Weitz, Lindsey Vanderlyn, Ngoc Thang Vu, and Elisabeth André. 2021. "It's our fault!": Insights Into Users' Understanding and Interaction With an Explanatory Collaborative Dialog System. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, Arianna Bisazza and Omri Abend (Eds.). Association for Computational Linguistics, Online, 1–16. doi:10.18653/v1/2021.conll-1.1

[141] John R. Wilson and Andrew Rutherford. 1989. Mental Models: Theory and Application in Human Factors. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 31, 6 (1989), 617–634. doi:10.1177/001872088903100601

[142] Julia Woodward, Zari McFadden, Nicole Shiver, Amir Ben-hayon, Jason C. Yip, and Lisa Anthony. 2018. Using Co-Design to Examine How Children Conceptualize Intelligent Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3174149

[143] Robert H. Wortham, Andreas Theodorou, and Joanna J. Bryson. 2017. Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 1424–1431. doi:10.1109/ROMAN.2017.8172491

[144] Yaqi Xie, Indu P Bodala, Desmond C. Ong, David Hsu, and Harold Soh. 2020. Robot capability and intention in trust-based decisions across tasks. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction* (Daegu, Republic of Korea) *(HRI '19)*. IEEE Press, 39–47.

[145] Qing Ru Yong and Alex Mitchell. 2023. From Playing the Story to Gaming the System: Repeat Experiences of a Large Language Model-Based Interactive Story. In *Interactive Storytelling*, Lissa Holloway-Attaway and John T. Murray (Eds.). Springer Nature Switzerland, Cham, 395–409.

[146] Richard M. Young. 2014. Surrogates and mappings: Two kinds of conceptual models for interactive devices. In *Mental models*. Psychology Press, 35–52. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315802725-4/surrogates-mappings-two-kinds-conceptual-models-interactive-devices-richard-young

[147] Tingru Zhang, Jiaqian Li, Linwei Qiao, Yong Zhang, Weitao Li, and Siu Shing Man. 2025. Evaluation of Drivers' Mental Model, Trust, and Reliance Toward

Level 2 Automated Vehicles. *International Journal of Human–Computer Interaction* 41, 6 (2025), 3696–3707. doi:10.1080/10447318.2024.2412422

[148] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 156, 26 pages. doi:10.1145/3613904.3642385

A Search keywords

Table 2 shows the search keywords used in each digital library.

B Co-occurrences of the type(s) of mental model assumed (epistemic assumptions) across the type of AI system considered in a paper

Figure 9 depicts the paper-level distribution of epistemic assumptions underlying mental model elicitation across AI system clusters.

Digital library	Adaptation	Search query
ACM Digital Library	X	`Abstract: ("artificial intelligence" OR "AI" OR "machine learning" OR "ML" OR system* OR agent* OR algorithm* OR automat*) AND ("mental model" OR user*)` `Full text: "mental model" AND "participants"`
IEEE Xplore	X	`Abstract: ("artificial intelligence" OR "AI" OR "machine learning" OR "ML" OR system* OR agent* OR algorithm* OR automat*) AND ("mental model" OR user*)` `Full text only: "mental model" AND "participants"`
OpenAlex	X	`Title and abstract: ("artificial intelligence" OR " ML " OR " AI " OR "machine learning" OR "systems" OR agent OR algorithm* OR automat*) AND "mental model"` `Full text: "mental model" AND "participants"`
Springer Link	Search in abstract and full text not supported. Only extracted the 1000 most relevant results (export limit).	`Keywords: ("artificial intelligence" OR " ML " OR " AI " OR "machine learning" OR "systems" OR agent OR algorithm* OR automat*) AND "mental model"` Searched in articles, research articles, and conference papers only.
Google Scholar	Distinct search in abstract or full text not supported.	`Find articles with all of the words: "mental model"` `at least one of the words: "artificial intelligence" AI "machine learning" ML`

Table 2: Search keywords used for each digital library.

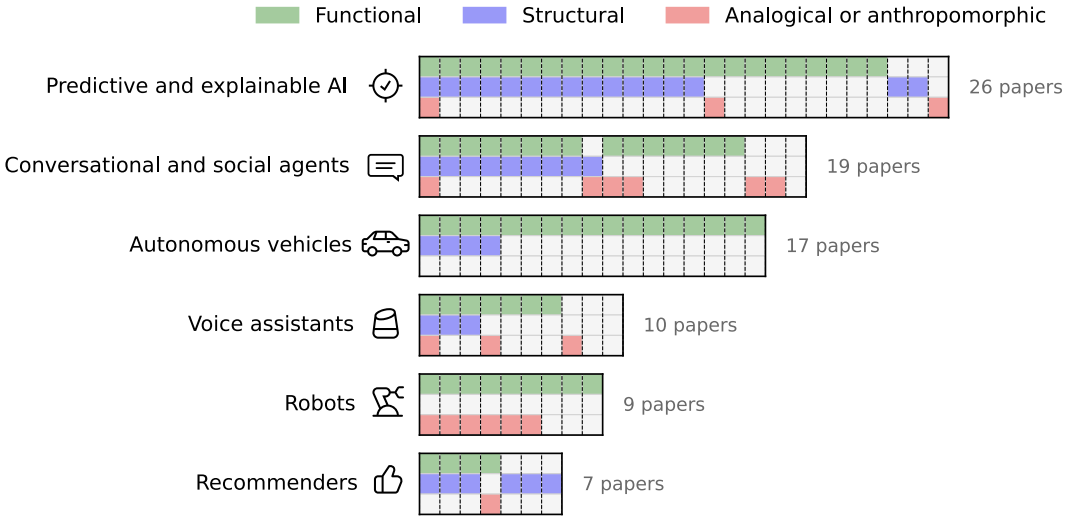


Figure 9: Paper-level co-occurrences of the type mental model assumed (epistemic assumption) across AI system clusters. Each column in a table represents one paper, and colored cells indicate the epistemic assumptions underlying the mental model elicitation described in that paper. Columns without colored cells indicate unclear or unspecified epistemic assumptions according to our coding scheme.